

Benchmarking machine learning against a regulatory rule-based baseline for drinking-water compliance classification

Aljaž Sovič

University of Ljubljana

Barbara Hribar Lee

University of Ljubljana

Zoran Bosnić

`zoran.bosnic@fri.uni-lj.si`

University of Ljubljana

Article

Keywords: machine learning, drinking water quality, compliance classification, rule-based baseline, SHAP, metolachlor

Posted Date: October 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11142900/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Benchmarking machine learning against a regulatory rule-based baseline for drinking-water compliance classification

Aljaž Sovič¹, Barbara Hribar Lee², Zoran Bosnić^{1*}

^{1*}Faculty of Computer and Information Science, University of
Ljubljana, Večna pot 113, Ljubljana, 1000, Slovenia.

²Faculty of Chemistry and Chemical Technology, University of
Ljubljana, Večna pot 113, Ljubljana, 1000, Slovenia.

*Corresponding author(s). E-mail(s): zoran.bosnic@fri.uni-lj.si;
Contributing authors: as67643@student.uni-lj.si;
barbara.hribar@fkkt.uni-lj.si;

Abstract

Machine learning is increasingly proposed for environmental compliance assessment, yet it is rarely benchmarked against the regulatory rule set it would replace. We formulate drinking-water compliance as binary classification and compare a rule-based regulatory baseline with four machine-learning models (gradient boosting, random forest, LightGBM, multi-layer perceptron) on 22,152 national monitoring samples spanning 2019–2025, using Leave-One-Year-Out cross-validation and SHAP-based feature selection. The learned models match but never exceed the baseline: domain-defined thresholds already capture almost all of the available signal. Two domain interventions prove more valuable than model choice, namely a regulatory correction for metolachlor degradation products in the pesticide-sum rule, which more than halves baseline errors from 19 to 9, and a two-step consensus model for rare violations. We conclude that a rule-based baseline is a necessary control in regulatory monitoring, and that the protocol transfers to other threshold-governed environmental domains.

Keywords: machine learning, drinking water quality, compliance classification, rule-based baseline, SHAP, metolachlor

Acknowledgements

This work was carried out within the framework of the Africa-Europe Clusters of Research Excellence (CoRE) project Water Resources Management for a Sustainable and Just Future. B.H.-L. and Z.B. acknowledge the support of the Water Research Center at the University of Ljubljana and the University of Ljubljana Development Fund (RSUL). The authors are grateful to dr. Brigita Jamnik at VOKA Snaga for her valuable assistance in clarifying and interpreting data used in this study.

1 Introduction

Drinking-water quality is a fundamental determinant of public health and of the reliable functioning of modern society. Despite advanced treatment and distribution processes, monitoring the compliance of individual samples remains a continuous obligation of water-supply systems, because even short-term exceedances of individual indicators can have serious health consequences. In Slovenia, a national monitoring programme systematically regulates drinking-water quality, and within it thousands of samples are collected and analysed each year from water-supply systems across the entire country. In current practice, the compliance of a sample is determined by directly comparing the measured values against the limit values prescribed by the Rules on Drinking Water ([Ministrstvo za zdravje, 2004](#)) and the Decree on Drinking Water ([Vlada Republike Slovenije, 2023](#)), which constitutes a deterministic, domain-grounded, and at the same time the only established assessment method.

The problem addressed in this study is the prediction of drinking-water sample compliance with the prescribed quality indicators, based on measured physical, chemical, and microbiological parameters. The task is formulated as a binary classification problem within supervised machine learning, in which each sample is assigned to one of two classes: compliant or non-compliant. Compliance in this context denotes the sample's fitness for human consumption, whereas non-compliance means that at least

one measured parameter exceeds its prescribed limit value. In the dataset considered here, comprising 22,152 samples from the 2019–2025 period, only about 9.4% of samples are non-compliant, which makes the classification task strongly imbalanced and further complicates the use of standard machine-learning approaches.

The relevance of the problem is twofold. First, machine-learning methods could capture complex patterns among parameters that a deterministic threshold-checking approach cannot, potentially enabling a more comprehensive understanding of the drivers of non-compliance and earlier warning of at-risk samples. Second, although machine learning is well established for comparable environmental problems (Hu, Dai, Sun, & Sunderland, 2023), published applications are seldom benchmarked against the regulatory rule set they are intended to improve upon, so it remains unclear how much predictive value learned models add over the deterministic thresholds already in force. Multi-year national monitoring archives, of which the data used here are one instance, offer the scale and the label provenance needed to close that gap.

Several predictive approaches are evaluated. As a reference baseline, a rule-based (RB) approach is adopted that directly emulates current practice by checking exceedances of the prescribed limit values. As representative machine-learning models, Gradient Boosting (GB), Random Forest (RF), LightGBM, and a multi-layer perceptron (MLP) are selected. The most influential features are selected with SHAP (SHapley Additive exPlanations), and predictive performance is assessed using Leave-One-Year-Out (LOYO) time-series cross-validation, in which each iteration uses one year as the test set and the remaining years as the training set. Beyond these established approaches, two additional methods are proposed. The first is a metolachlor regulatory correction, a preprocessing step that removes a time-dependent labelling inconsistency in the dataset arising from the tightening of regulatory criteria for the pesticide metabolites metolachlor ESA and metolachlor OXA in 2025. The second is a two-step model that combines the rule-based approach and a machine-learning model

139 by requiring the agreement of both, thereby aiming to exploit any complementarity
140 in their errors.
141

142 The main objective of this work is a systematic comparison of these approaches and
143 an assessment of whether machine-learning approaches surpass the rule-based baseline
144 for the problem at hand. Within this objective, both proposed methods are evaluated
145 specifically: the effect of the metolachlor regulatory correction on the performance
146 of the individual approaches, and whether the two-step model successfully exploits
147 the differences between the two approaches. The contribution of derived features, in
148 particular the total pesticide concentration sum, is additionally assessed. Results are
149 reported separately for the base scenario and for the scenario after the regulatory
150 correction, providing insight into the interplay between the individual interventions.
151

152 Because the comparison is anchored in the regulatory rules rather than in any
153 particular parameter set, the resulting protocol is not specific to drinking water. Any
154 monitoring regime in which compliance is defined by thresholds on measured param-
155 eters – ambient air quality, wastewater discharge limits, soil contamination screening
156 values – admits the same three-part evaluation: a rule-based baseline reconstructed
157 from the legal text, learned models restricted to the same inputs, and a temporal val-
158 idation split that respects the dates on which the rules changed. The two proposed
159 interventions generalise in the same way: the metolachlor correction is an instance of
160 repairing labels invalidated by a mid-series regulatory amendment, and the two-step
161 consensus model is an instance of testing whether a learned model and a codified rule
162 set make complementary errors.
163

164 The remainder of this paper is organised as follows. Section 2 reviews existing
165 approaches to water-quality prediction, from classical statistical methods to modern
166 classification and regression machine-learning models. Section 3 describes the dataset,
167 feature engineering, SHAP-based feature selection, the predictive models, hyperparam-
168 eter selection, the validation protocol, and the evaluation metrics. Section 4 presents
169

the results of the systematic comparison across scenarios, the error analysis, and a discussion of the findings. Section 5 summarises the main findings, outlines the limitations of the work, and indicates directions for future research.

2 Related work

A range of methodological approaches to the prediction and assessment of drinking-water quality has become established in the literature. Traditional statistical methods provide a baseline and a reference framework for comparison with more recent solutions, whereas over the past decade machine-learning models (Hu et al., 2023) have moved to the fore; these are commonly divided into classification and regression approaches.

Among the traditional statistical approaches, Maiolo and Pantusa (2021) analysed drinking-water quality in the Emilia-Romagna region of northern Italy: using principal component analysis (PCA) and k -means clustering over a set of 18 physico-chemical parameters, they resolved the temporal and spatial structure of water quality for the 2013–2020 period, finding that only two components jointly explained almost 90% of the variance, one associated with mineral content (Ca^{2+} , Mg^{2+} , hardness) and the other with nutrients. Katimon, Shahid, and Mohsenipour (2018) developed stochastic ARIMA time-series models for the Johor River system in Malaysia, analysing drinking-water quality (pH, colour, turbidity, iron, and others) together with hydrological variables (precipitation and river discharge) for 2004–2007; the ARIMA model proved simple and robust for forecasting the quality of water resources without resorting to more complex machine-learning methods.

Classification models are most often aimed at determining the compliance of samples with legal standards, either as binary classification (potable versus non-potable water) or as multi-class ranking by quality class; they are widely used because of the interpretive simplicity of their outputs and their direct usefulness for decision-making.

231 Alongside conventional machine-learning models, alternative methods such as fuzzy-
 232 logic systems also address the water-quality classification problem, with the literature
 233 consistently reporting high accuracy. [Chen, Xu, Pan, Zhao, and Song \(2024\)](#) achieved
 234 96.2% accuracy on a publicly available set of about 200,000 groundwater samples using
 235 a hybrid BS-MLP method (gradient boosting combined with a multi-layer perceptron),
 236 indicating the potential of hybrid approaches for classifying complex environmental
 237 data. [Ahmed et al. \(2019\)](#) classified a discrete water-quality class (WQC, a nominal
 238 scale of the water quality index, WQI) with an MLP neural network, reaching 85%
 239 accuracy. [Tadesse and Dinka \(2023\)](#) tested k -nearest neighbours and a decision tree
 240 for classifying reservoir water quality in South Africa on the basis of 40 years of data,
 241 with both models attaining 96.4% and 96.2% accuracy, respectively, when assigning
 242 samples to three quality classes (poor, medium, good). [Barzegar, Gorelova, Bellini,
 243 and D’Ascenzo \(2023\)](#) used a hierarchical Mamdani fuzzy inference system (FIS), com-
 244 posed of three intermediate stages and one final stage that jointly account for nine
 245 water-quality parameters, to assess drinking-water quality in Rome, and reported that
 246 fuzzy-logic systems perform effectively in this domain. [Hu et al. \(2023\)](#) reviewed 27
 247 studies published between 2012 and 2022, in which Random Forest models with binary
 248 classification of limit-value exceedances were most frequently developed for predicting
 249 various drinking-water quality parameters; these models achieved high classification
 250 accuracies and reliably flagged at-risk areas of contaminated water. [Farzana, Paudyal,
 251 Chadalavada, and Alam \(2023\)](#) compared several machine-learning and deep-learning
 252 models for predicting the WQI in three reservoirs in Australia over 2000–2022, find-
 253 ing that XGBoost (extreme gradient boosting) and a gated recurrent unit (GRU)
 254 neural network attained the highest accuracy and successfully captured changes in
 255 precipitation patterns.

272 Regression models are intended for predicting continuous values of water parame-
 273 ters, such as pH, turbidity, or individual substance concentrations, and are therefore
 274
 275
 276

particularly suited to analysing temporal trends and quantifying deviations from expected values; a direct comparison of regression and classification models on the basis of reported metrics is not possible, as the former rely on continuous predictions (MAE, RMSE) and the latter on classification accuracy. [Im, Song, Lee, and Cho \(2022\)](#) developed deep neural networks for a 24-hour forecast of tap-water parameters (pH, turbidity, and chlorine content) from five years (2017–2021) of data from 33 treatment plants in South Korea; long short-term memory (LSTM), GRU, and the SCINet (sample convolution and interaction network) architecture were trained on sliding three-day windows, and an optimised combined model reached an average accuracy of 98.8% with a mean absolute error (MAE) of approximately 0.0003. [Ahmed et al. \(2019\)](#) also predicted the WQI from basic water parameters (pH, chlorine, nitrites, total dissolved solids, and others) using gradient boosting and polynomial regression, with both models achieving an MAE below 3. [Costa et al. \(2024\)](#) tested decision trees, Random Forest, Gradient Boosting, and XGBoost for predicting biochemical oxygen demand (BOD), nitrate, and phosphate concentrations in Brazilian rivers, with XGBoost attaining the lowest root mean square error (RMSE) for all three target variables.

Overall, the literature shows that research over the past decade has concentrated strongly on the use of machine learning for drinking-water quality prediction. Classification models such as Random Forest, XGBoost, support vector machines, k -nearest neighbours, and neural networks reach accuracies above 95% in most of the studies considered ([Chen et al., 2024](#); [Tadesse & Dinka, 2023](#)), enabling reliable identification of at-risk areas, whereas regression models, although less accurate, predict continuous parameter values useful for time-series analysis and trend monitoring ([Ahmed et al., 2019](#); [Im et al., 2022](#)). Classical statistical methods such as ARIMA and PCA provide a robust and interpretable baseline for comparison with modern approaches ([Katimon et al., 2018](#); [Maiolo & Pantusa, 2021](#)). Ensemble models, in particular Random Forest and Gradient Boosting, consistently achieve high classification accuracies

323 across diverse geographical and data contexts and prove robust when working with
324 real, imbalanced drinking-water quality datasets (Farzana et al., 2023; Hu et al., 2023);
325 these findings informed the design of the approach described in the following.
326
327

328 329 **3 Materials and methods** 330

331 332 **3.1 Problem formulation** 333

334 The prediction of drinking-water compliance is treated as a binary classification prob-
335 lem within supervised machine learning. For each sample $\mathbf{x} \in \mathcal{X}$, described by a vector
336 of n measured water-quality parameters, a function h is sought that maps the sample
337 to a compliance label:
338
339

$$340 \quad h : \mathcal{X} \rightarrow \{0, 1\}, \quad (1)$$

341 where $y = 1$ denotes a non-compliant sample, in which at least one measured value
342 exceeds its prescribed limit, and $y = 0$ a compliant sample. The classification task is
343 strongly imbalanced, as only about 9.4% of samples in the dataset are non-compliant.
344 For such problems, classification accuracy is an unsuitable primary performance mea-
345 sure, since a trivial model that labels all samples as compliant attains a high accuracy
346 yet has no practical value. Evaluation is therefore centred on the macro-averaged F_1
347 measure, which equalises the contribution of the minority and majority classes and is
348 defined in Section 3.8.
349
350
351
352
353
354
355

356 357 **3.2 Dataset** 358

359 The dataset was assembled from seven annual summary reports on drinking-water
360 quality monitoring in Slovenia for the 2019–2025 period (Ministrstvo za zdravje, 2025).
361 Each annual report is a separate Excel (xlsx) file containing the laboratory analysis
362 results of individual water samples taken from water-supply systems across the entire
363 country. The raw dataset comprises 22,152 samples and 231 measured parameters,
364
365
366
367
368

which were not uniformly represented across all years. In addition to the measured chemical, physical, and microbiological properties, metadata such as the year, month, region, water-source type, and type of water treatment are recorded for each sample.

Owing to the inconsistent representation of individual parameters across years and to undefined measurement units or limit values for some indicators, 62 measured parameters were retained from the initial set of 231 for further analysis, namely those for which appropriate units and prescribed limit values or sufficiently complete records were available. The retained parameters were grouped into five content categories: microbiological parameters (6 indicators, including the presence of *E. coli*, enterococci, and coliform bacteria), physical parameters (7 indicators, including turbidity, electrical conductivity, and pH), chemical parameters (13 indicators, including nitrate, ammonium, and trihalomethanes), metals (8 indicators, including lead, arsenic, and manganese), and pesticides (28 indicators). Together with six sample metadata fields, this yields 68 attributes in the base scenario and in the scenario with the metolachlor regulatory correction, and 69 attributes in the scenarios that include the derived total pesticide sum (Section 3.3). For 57 of the retained measured parameters, a prescribed limit value is available from the drinking-water regulations ([Ministrstvo za zdravje, 2004](#); [Vlada Republike Slovenije, 2023](#)), which the rule-based approach uses for direct classification.

In the laboratory records, values below the detection limit are often reported as strings (e.g. <0.007). Such records were converted to the numerical value corresponding to the respective detection limit, enabling numerical processing without introducing systematic bias. Missing values in non-microbiological parameters were imputed with the median, whereas the microbiological parameters were complete in the original data and required no imputation. As the distributions of some metals are strongly right-skewed and contain outliers that would disproportionately affect model

training, extreme values were capped at the 99th percentile of the respective distribution. After cleaning and preparation, the dataset used for modelling comprises 22,152 samples, of which 2,079 (9.4%) are labelled as non-compliant.

3.3 Feature engineering

Two feature-engineering interventions were applied to the input vector: the introduction of a derived variable and a preprocessing step that removes a regulatory inconsistency in the pesticide metabolites. One derived variable is included, namely the total pesticide concentration sum, which the drinking-water regulations define as a separate indicator with its own limit value. A set of binary violation flags, encoding threshold exceedance for each parameter with a prescribed limit, was also considered but not added as separate variables, since the machine-learning models access this information directly through the raw measured values.

For the pesticide metabolites metolachlor ESA and metolachlor OXA, a systematic labelling inconsistency across years was identified, the removal of which is one of the contributions of this work. Both metabolites were measured before 2025 but were not yet regulatory-relevant, so their exceedances did not affect the official compliance label; with the tightening of the criteria in 2025 they acquired regulatory validity, and values above the prescribed threshold from that year onward render a sample non-compliant ([Nacionalni inštitut za javno zdravje, 2025](#)). This time-dependent change produced an inconsistency in the labels: pre-2025 samples with values that would exceed the 2025 threshold were nonetheless labelled compliant, whereas 2025 samples with analogous values are labelled non-compliant. To remove this inconsistency, a preprocessing step referred to as the metolachlor regulatory correction was introduced, in which the values of both metabolites are replaced with a missing value for all samples from before 2025:

$$x_{\text{ESA}}^{(i)} \leftarrow \text{NaN}, \quad x_{\text{OXA}}^{(i)} \leftarrow \text{NaN}, \quad \text{if } t^{(i)} < 2025, \quad (2)$$

where $t^{(i)}$ denotes the sampling year of sample i . This converts 368 values to missing while retaining the 118 non-zero, regulatory-valid measurements from 2025; it does not alter the compliance labels but only removes the regulatory-invalid measurements from the input of both approach families. A narrower inconsistency remains within 2025 itself, and its effect is analysed in Section 4.8.

3.4 Feature selection with SHAP

For models with several dozen attributes, interpreting the contribution of individual parameters is non-trivial, while a model with a smaller attribute set is often more robust and less exposed to noise. The most important attributes were selected using SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), in which the contribution of each attribute to the prediction is evaluated through its Shapley value from coalitional game theory; the method satisfies the axioms of local accuracy, consistency, and missingness. For each attribute j , the mean absolute SHAP value over the whole dataset is computed:

$$\bar{\phi}_j = \frac{1}{N} \sum_{i=1}^N \left| \phi_j(\mathbf{x}^{(i)}) \right|, \quad (3)$$

where $\phi_j(\mathbf{x}^{(i)})$ denotes the SHAP contribution of attribute j for sample i and N is the number of samples. The attributes are then ranked in descending order of $\bar{\phi}_j$, with the attribute having the greatest average influence on the model prediction at the top.

To analyse the trade-off between the number of attributes and predictive performance systematically, a detailed evaluation was carried out at different values of K (a top- K sweep) for $K \in \{5, 10, 15, 20, 30, 40, K_{\max}\}$, where $K_{\max}$ equals the size of the full attribute set in the given scenario (68 in the base scenario and in the scenario with the metolachlor regulatory correction, and 69 in the scenarios that include the derived total pesticide sum). At each value of K , the models were trained and evaluated on the reduced set, yielding a curve of the F_1 measure as a function of K . This makes it possible to determine the minimum number of attributes that still provides predictive

507 performance comparable to the full set, and reveals any saturation at which adding
508 further attributes no longer improves, or even degrades, predictive performance.
509

510

511

512 **3.5 Predictive models**

513

514 All predictive approaches used in the study are described below. The rule-based ref-
515 erence approach, which serves as the comparison baseline, is presented first, followed
516 by the four machine-learning models, which are robust for imbalanced tabular clas-
517 sification (Section 2) and support the attribute-importance computation required for
518 SHAP-based selection. The two-step model, which combines the rule-based approach
519 and machine learning and is a further contribution of this work, is described last.
520

521

522

523 **3.5.1 Rule-based reference approach**

524

525 The rule-based reference approach directly emulates the legislator’s decision-making.
526
527 A sample is labelled non-compliant if at least one measured parameter exceeds its limit
528 value prescribed in the drinking-water regulations. Formally, the rule-based prediction
529 is given by
530

531

$$\hat{y}_{\text{RB}}(\mathbf{x}) = \begin{cases} 1, & \text{if } \exists j \in S : x_j > t_j, \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

533

534 where S is the subset of parameters with a known limit value and t_j the corresponding
535 threshold. The approach is deterministic, fully interpretable, and requires no training,
536 which makes it a natural baseline for comparison with machine-learning models.
537

538

539

540 **3.5.2 Machine-learning models**

541

542 Four machine-learning models were evaluated. *Gradient Boosting* (Friedman, 2001)
543 builds a strong predictor by combining a sequence of weak learners, typically shallow
544 decision trees. At step m , a new tree $h_m(\mathbf{x})$ is fitted to the pseudo-residuals r_{im} ,
545 computed as the negative gradient of the chosen loss function L with respect to the
546

547

548

current model prediction F_{m-1} :

$$r_{im} = - \left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)} \right]_{F=F_{m-1}}, \quad (5)$$

and the prediction is updated as $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu h_m(\mathbf{x})$, where $\nu \in (0, 1]$ is the learning rate. *Random Forest* (Breiman, 2001) aggregates the predictions of many independently trained decision trees; each tree is trained on a bootstrap sample and considers only a random subset of attributes at each split, which decorrelates the trees and, through averaging, prevents the overfitting characteristic of individual deep trees, while remaining relatively robust to outliers and partially missing values. *LightGBM* (Ke et al., 2017) is a gradient-boosting implementation that introduces two key optimisations: Gradient-based One-Side Sampling (GOSS), which retains high-gradient instances and sub-samples low-gradient ones, and Exclusive Feature Bundling (EFB), which merges mutually exclusive attributes to reduce dimensionality; unlike classical gradient boosting, it grows trees leaf-wise rather than level-wise, which speeds up loss reduction but requires more careful depth tuning to limit overfitting. Finally, the *multi-layer perceptron* (MLP) (Rumelhart, Hinton, & Williams, 1986) is a fully connected neural network with one or more hidden layers, in which a sequence of linear maps interleaved with non-linear activations is followed by a sigmoid output that estimates the probability of the positive class; it was trained by minimising the binary cross-entropy with the Adam optimiser (Kingma & Ba, 2015). The MLP is included primarily as a control, since on tabular data it often does not surpass tree-based methods but offers a different inductive bias and enables comparison between the two model classes.

3.5.3 Two-step model

The two-step model is designed as a further contribution that exploits the complementarity of the two approach families, the rule-based and the machine-learning

one. In this design, a sample is labelled non-compliant only when both approaches independently identify it as such. Formally, the two-step prediction is given by

$$\hat{y}_{2\text{step}}(\mathbf{x}) = \hat{y}_{\text{RB}}(\mathbf{x}) \wedge \hat{y}_{\text{ML}}(\mathbf{x}), \quad (6)$$

where $\wedge$ denotes logical conjunction. The operating scheme is shown in Figure 1.

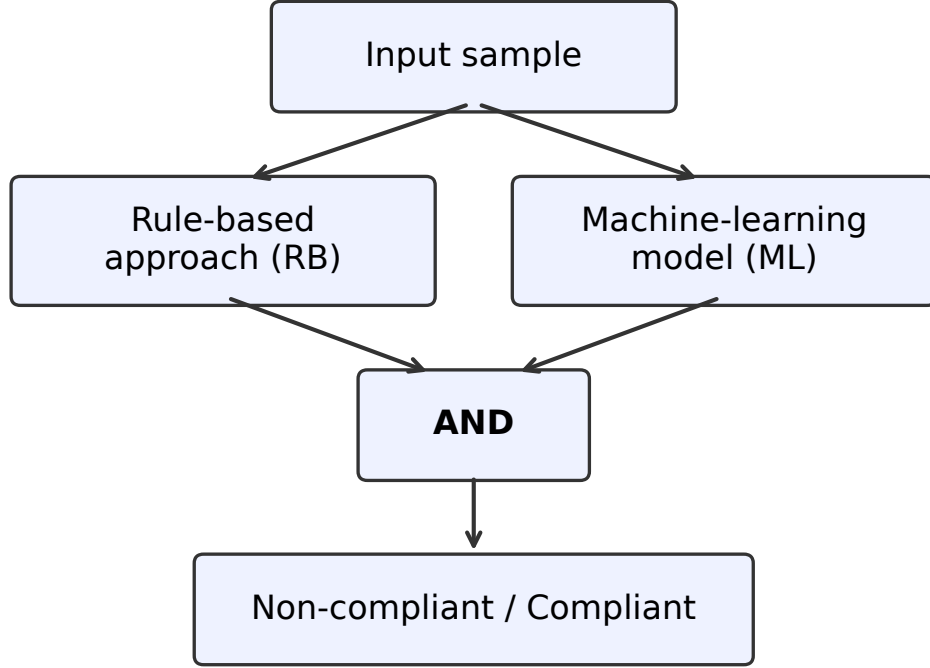


Figure 1: Scheme of the two-step model. An input sample is processed in parallel by the rule-based approach and the machine-learning model, and is labelled non-compliant only if both branches return a positive prediction.

The design rests on the observation that the two approach families typically err on different samples: the rule-based approach is sensitive to labelling inconsistencies and regulatory changes, whereas the machine-learning models may fail on rare samples

whose violations lie outside the set of top-ranked attributes. Requiring the agreement of both approaches means that a false positive arises only when the error occurs in both branches simultaneously, which in practice substantially lowers the number of false positives; conversely, the consensus requirement may slightly increase the number of false negatives, since it suffices for one approach to miss a non-compliant sample for the combined prediction to miss it as well. Such a trade-off is reasonable when the consequences of false positives (for example, unnecessary alerts or the additional cost of repeat analyses) outweigh those of false negatives, and both components are already highly sensitive on their own. This trade-off is analysed in detail in Section 4.7.

3.6 Hyperparameter selection

All machine-learning models have hyperparameters, such as the number of trees, maximum depth, learning rate, or network architecture, that must be set before training. These were selected with the Optuna framework (Akiba, Sano, Yanase, Ohta, & Koyama, 2019), which formulates the hyperparameter search as a Bayesian optimisation process using the Tree-structured Parzen Estimator (TPE) to propose candidates. The optimisation criterion was the macro-averaged F_1 measure (Section 3.8), computed as the average over 5-fold cross-validation on the training set. For each model, 30 trials were run, which was sufficient for convergence of the estimated objective in preliminary experiments. The selected hyperparameters were then used in the final evaluation within the validation procedure described in Section 3.7.

3.7 Validation protocol

A standard random split of the dataset into training and test sets (random holdout) overestimates predictive performance for time-ordered data, since samples in the two sets share the calendar year and often the geographical location. As the model in practical use predicts for samples from years not yet represented in the training set,

691 a methodological decision was taken to use Leave-One-Year-Out (LOYO) time-series
692 cross-validation instead of a random split. The cross-validation scheme itself is an
693 established protocol, but its adaptation to a split by sampling year follows directly
694 from the nature of the dataset. In the LOYO procedure, the data are divided into
695 seven folds, each corresponding to one year in the dataset; in iteration i , year i is used
696 as the test set and the remaining six years as the training set. The scheme is shown
697 in Figure 2. This design directly emulates the practical use case, in which the training
698 data come from past years and the model predicts the compliance of samples in the
699 following year. The final performance estimate is the average of the F_1 measure over
700 the seven folds, and the standard deviation is used as a measure of the stability of
701 predictions across years.
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736

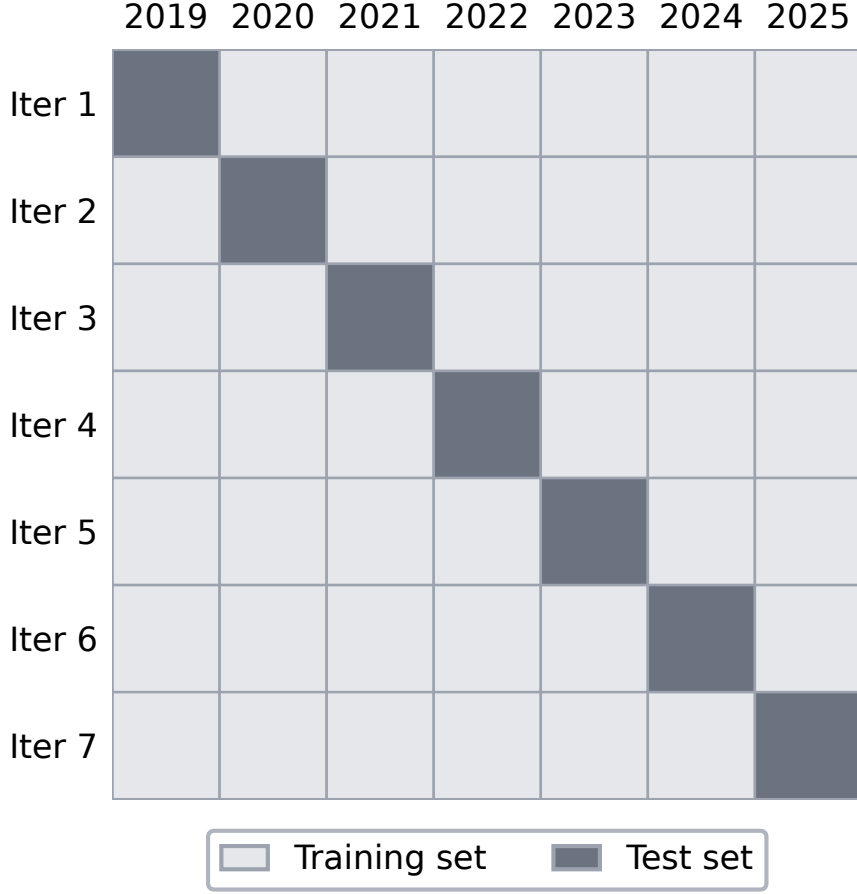


Figure 2: Scheme of the LOYO time-series cross-validation. In each iteration, one year is used as the test set (dark cell) and the remaining six as the training set (light cells). The procedure is repeated for all years in the dataset.

3.8 Evaluation metrics and statistical tests

For a strongly imbalanced classification problem, accuracy is an unsuitable central performance measure, since a trivial model that classifies all samples as compliant attains about 90.6% accuracy yet has no practical value. The primary measure was therefore the macro-averaged F_1 measure, computed as the unweighted average of the F_1 measure across both classes (compliant and non-compliant). For a strongly

783 imbalanced class distribution, such averaging equalises the contribution of the minority
784 and majority classes, preventing a high value on the majority class from masking poor
785 performance on the minority class. The F_1 measure for an individual class is defined
786 as the harmonic mean of precision and recall:
787

$$789 \text{ Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (7)$$

$$793 F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (8)$$

796 where TP, FP, and FN denote the number of true-positive, false-positive, and
797 false-negative predictions for the class considered. Precision and recall are reported
798 separately in addition to the overall F_1 measure, likewise as macro averages, as they
799 allow interpretation of the nature of each model’s errors. All performance measures
800 are computed within the LOYO procedure (Section 3.7) separately for each annual
801 fold, and the final reported value is their average over the seven years.

802 To compare two models on the same test set, McNemar’s test for paired categorical
803 data (McNemar, 1947) was used, which tests the null hypothesis that the two mod-
804 els are equally accurate on the basis of the samples on which their predictions differ.
805 To assess the uncertainty of the difference in the F_1 measure, a bootstrap procedure
806 (Efron, 1979) with 10,000 repetitions was additionally carried out and the 95% confi-
807 dence interval of the difference computed. A difference between two models is treated
808 as statistically significant when the confidence interval does not contain zero and, at
809 the same time, $p < 0.05$ in McNemar’s test. To compare performance at the level of
810 individual cross-validation folds, a paired t -test was additionally used, comparing the
811 pairs of F_1 values across the seven LOYO folds between two approaches; the test eval-
812 uates the null hypothesis that the mean of the differences $F_1^A - F_1^B$ across the seven
813 folds equals zero, and detects systematic performance differences that are consistent
814 across all years in the dataset.

4 Results and discussion	829
	830
4.1 Experimental design and research questions	831
	832
	833
The empirical evaluation is organised around four research questions that follow	834
directly from the aims of the work and the choice of approaches in Section 3:	835
	836
	837
• RQ1: Do machine-learning models surpass the rule-based reference approach on	838
the task considered?	839
	840
• RQ2: Do derived features, in particular the total pesticide concentration sum,	841
contribute to predictive performance?	842
	843
• RQ3: Does the proposed metolachlor regulatory correction improve the perfor-	844
mance of the individual approaches, and how does it affect their relationship?	845
	846
• RQ4: Does the proposed two-step model exploit the complementarity of the two	847
approach families and achieve better results than either of them alone?	848
	849
	850
	851
To answer these questions, three scenarios were tested that differ systematically	852
in the presence or absence of the individual data interventions. The base scenario	853
is a direct comparison of the rule-based approach with the machine-learning mod-	854
els, without derived features and without the metolachlor regulatory correction. The	855
second scenario adds the total pesticide concentration sum as an additional derived fea-	856
ture. The third scenario introduces the metolachlor regulatory correction described in	857
Section 3.3. In each scenario, the two-step model (Section 3.5.3) is evaluated alongside	858
the rule-based approach and the standalone machine-learning models. All scenarios use	859
the same evaluation procedure: LOYO time-series cross-validation (Section 3.7) over	860
2019–2025, giving seven independent performance estimates whose mean and stan-	861
dard deviation are used for comparison. For each approach, performance is evaluated	862
at different attribute-set sizes $K \in \{5, 10, 15, 20, 30, 40, K_{\max}\}$, with the top-ranked K	863
attributes selected by SHAP (Section 3.4), and the hyperparameters are determined	864
per scenario with Optuna (Section 3.6).	865
	866
	867
	868
	869
	870
	871
	872
	873
	874

Rank	Attribute	$\bar{\phi}_j$	Category
1	coliform bacteria	11.52	microbiological
2	colony count at 22 °C	3.79	microbiological
3	enterococci	2.48	microbiological
4	colony count at 37 °C	2.40	microbiological
5	<i>Clostridium perfringens</i>	1.25	microbiological
6	<i>Escherichia coli</i>	1.03	microbiological
7	turbidity	0.55	physical
8	pH	0.31	physical
9	residual free chlorine	0.25	chemical
10	water temperature	0.10	physical

Table 1 The ten top-ranked attributes by mean absolute SHAP value $\bar{\phi}_j$, computed on a Gradient Boosting model in the base scenario.

The SHAP analysis presented in Section 4.2 reveals an important methodological consequence for the evaluation. As the signal in the data is strongly concentrated in a few microbiological parameters whose exceedance is already a sufficient condition for non-compliance, the rule-based approach is expected to be relatively hard to surpass: only a small amount of additional signal remains that the rules do not exploit directly.

4.2 SHAP feature ranking

Before comparing the approaches, the SHAP attribute ranking that determines which attributes are used at each value of K is presented. The ranking was computed on a Gradient Boosting model trained on the full training set of the base scenario, with the importance of each attribute expressed by the mean absolute value $\bar{\phi}_j$ defined in Section 3.4. The ten top-ranked attributes are given in Table 1, and an extended view of the fifteen top-ranked attributes in Figure 3.

Three notable characteristics emerge from the ranking. First, the six top-ranked attributes belong to the microbiological group, confirming the observation from the literature (Section 2) that microbiological parameters are the central indicator of non-compliance in drinking water. Second, the difference in importance between the first attribute ($\bar{\phi}_1 = 11.52$) and the seventh ($\bar{\phi}_7 = 0.55$) is of the order of $20\times$, indicating a

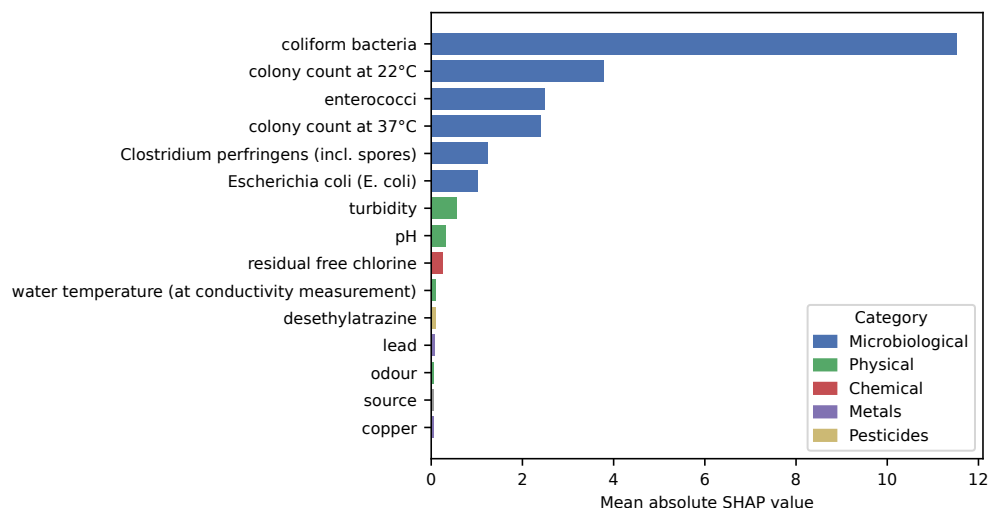


Fig. 3 Ranking of the fifteen top-ranked attributes by mean absolute SHAP value. Bars are coloured by parameter category (microbiological, physical, chemical, metals, pesticides).

strong concentration of the signal in the upper part of the ranking. Third, the first non-microbiological attribute (turbidity) appears only in seventh place, ranked just below the lowest-ranked microbiological attribute (*Escherichia coli*). The pesticide category as a whole appears only at lower ranks (ranks > 20), a consequence of the rarity of positive measurements and the correspondingly weak signal for these parameters.

The distribution of SHAP values across individual samples is shown in Figure 4, which reveals that the largest part of the prediction is driven by the above-average values of the top-ranked microbiological indicators. In routine drinking-water monitoring, where most samples are within limits for these indicators, the model therefore decides mainly on rare deviations, as is also evident from the strongly asymmetric shape of the SHAP value distribution.

4.3 Model comparison in the base scenario

This subsection compares the rule-based approach with the four machine-learning models (Gradient Boosting, Random Forest, LightGBM, and the multi-layer perceptron) in the base scenario, that is, without derived features and without the

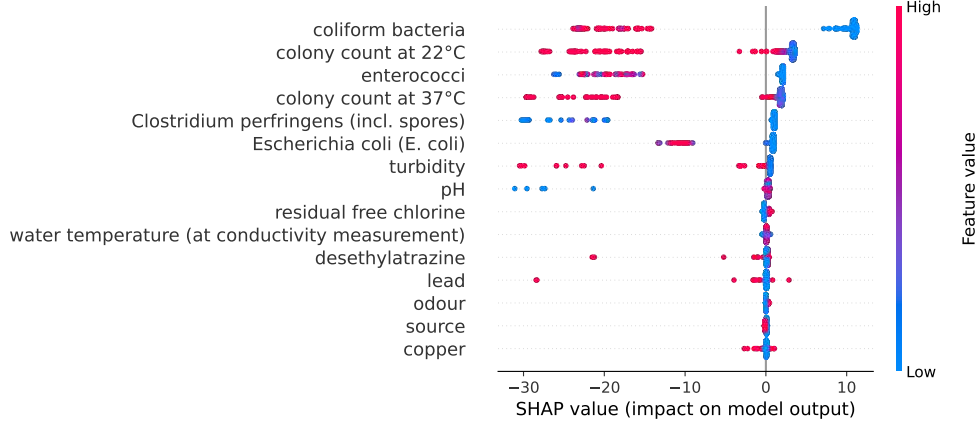


Fig. 4 Distribution of SHAP values across individual samples for the top-ranked attributes (beeswarm plot). Each point is one sample, and its position on the x -axis is its SHAP contribution to the non-compliance prediction. The point colour encodes the measured attribute value (red = high, blue = low). The attributes are ordered from the most to the least influential, and a point to the right of the centre increases the predicted probability of non-compliance while a point to the left decreases it. The top microbiological indicators (coliform bacteria, colony counts, enterococci) show by far the widest spread and thus the strongest effect, with their elevated values driving samples towards the non-compliant class.

K	GB	RF	LightGBM	MLP	RB
5	0.986 ± 0.003	0.986 ± 0.003	0.986 ± 0.002	n/a	0.987 ± 0.003
10	0.994 ± 0.004	0.993 ± 0.002	0.995 ± 0.003	0.989 ± 0.004	0.995 ± 0.003
15	0.997 ± 0.002	0.995 ± 0.003	0.997 ± 0.003	n/a	0.997 ± 0.002
20	0.997 ± 0.002	0.994 ± 0.003	0.997 ± 0.003	n/a	0.997 ± 0.002
30	0.997 ± 0.002	0.994 ± 0.002	0.997 ± 0.003	n/a	0.998 ± 0.002
40	0.997 ± 0.002	0.994 ± 0.002	0.997 ± 0.003	n/a	0.996 ± 0.003
68	0.997 ± 0.002	0.994 ± 0.002	0.997 ± 0.003	0.966 ± 0.006	0.997 ± 0.003

Table 2 F_1 measure in the base scenario at selected values of K (mean $\pm$ standard deviation over the seven LOYO folds, 2019–2025). The highest value in each column is in bold. The MLP was evaluated only at $K = 10$ and $K = 68$, since it is not competitive on this problem; intermediate values are omitted. Missing values are denoted n/a.

metolachlor regulatory correction; the results directly address RQ1. The mean F_1 measure and its standard deviation over the seven folds are summarised in Table 2, and the performance curves as a function of K are shown in Figure 5.

The rule-based approach attains its highest F_1 measure at $K = 30$, namely 0.998 ± 0.002 , and declines slightly at larger K ; this non-monotonic behaviour arises

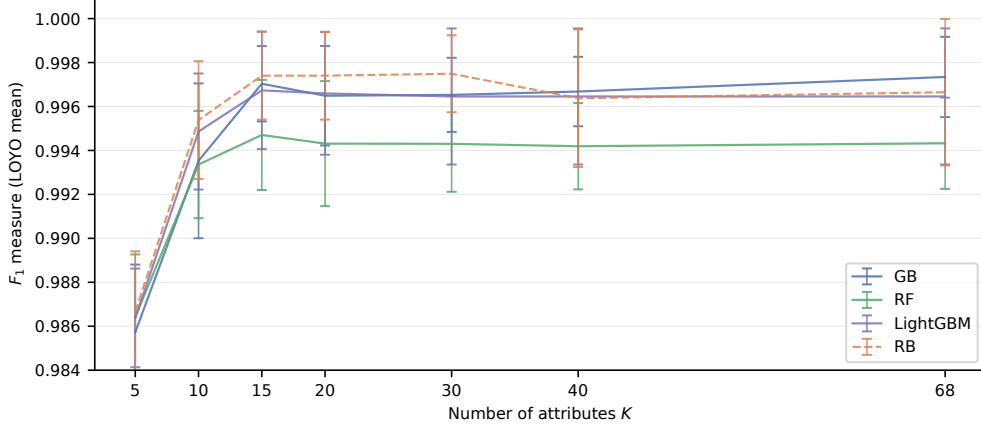


Fig. 5 F_1 measure as a function of the number of attributes K in the base scenario. Curves are the mean over the seven LOYO folds and the whiskers the standard deviation. The rule-based approach is shown with a dashed line. The multi-layer perceptron is not shown: it was evaluated only at $K = 10$ and $K = 68$ (Table 2), and its F_1 at $K = 68$ (0.966) lies below the displayed range.

because adding rarely measured attributes, particularly pesticides, to the rule set generates a limited number of false positives that slightly lower the F_1 measure. Among the machine-learning models, Gradient Boosting and LightGBM reach 0.997 at $K = 20$, matching the rule-based approach at the reported precision, with no statistically significant difference between them. At $K = 68$, Gradient Boosting attains the highest value among the machine-learning models (0.997), yet the difference from the rule-based approach remains within one standard deviation. Random Forest does not approach the rule-based approach at any value of K , and the multi-layer perceptron suffers a marked drop at the full attribute set ($F_1 = 0.966$), consistent with the weakness of dense neural networks on tabular data with a large proportion of rare attributes. A notable finding is that the rule-based approach reaches its maximum already at $K = 30$, meaning that defining rules for only the thirty top-ranked attributes suffices to exploit all available information; this is consistent with the SHAP analysis (Section 4.2). In absolute terms, the best base-scenario result (the rule-based approach

		$K = 30$			$K = 68$		
	Approach	Prec.	Recall	F_1	Prec.	Recall	F_1
	GB	0.998	0.995	0.997	0.999	0.996	0.997
	RF	0.998	0.990	0.994	0.999	0.990	0.994
	LightGBM	0.999	0.994	0.997	0.999	0.994	0.997
	RB	0.999	0.996	0.998	0.995	0.999	0.997

Table 3 Detailed performance measures (precision, recall, F_1) in the base scenario at $K = 30$ and $K = 68$, averaged over the seven LOYO folds. The MLP is omitted, as it was evaluated only at $K = 68$, where it attains $F_1 = 0.966$.

at $K = 30$) corresponds to 19 errors over 22,152 samples, which serves as the base-line error count for later comparison with the scenarios that include the metolachlor regulatory correction (Section 4.6).

For a detailed interpretation, Table 3 reports precision and recall of all approaches at $K = 30$, the best point of the rule-based approach, and at $K = 68$, the full attribute set.

The comparison of precision and recall suggests that the two approach families may make different kinds of errors. The machine-learning models achieve slightly higher precision (around 0.999), indicating that they less often label a compliant sample as non-compliant, but have slightly lower recall (up to 0.996). The rule-based approach shows the opposite tendency: lower precision (0.995) but higher recall (0.999). The overall F_1 measure is comparable across the three main approaches (GB, LightGBM, RB). As the absolute differences are small, confirming these observations would require statistical tests and/or a larger dataset.

4.4 Error analysis

This subsection compares the errors made by the individual approaches and examines performance at low K , in order to determine whether the errors of the machine-learning models and the rule-based approach overlap or are complementary. The value $K = 10$ was chosen because the difference between models is most informative there. At $K = 10$, the best machine-learning model (LightGBM) misclassifies 37 samples over

	LightGBM correct	LightGBM incorrect
RB correct	22,115	3
RB incorrect	0	34

Table 4 Distribution of errors at $K = 10$ between LightGBM and the rule-based approach. Values are the number of samples in each prediction combination.

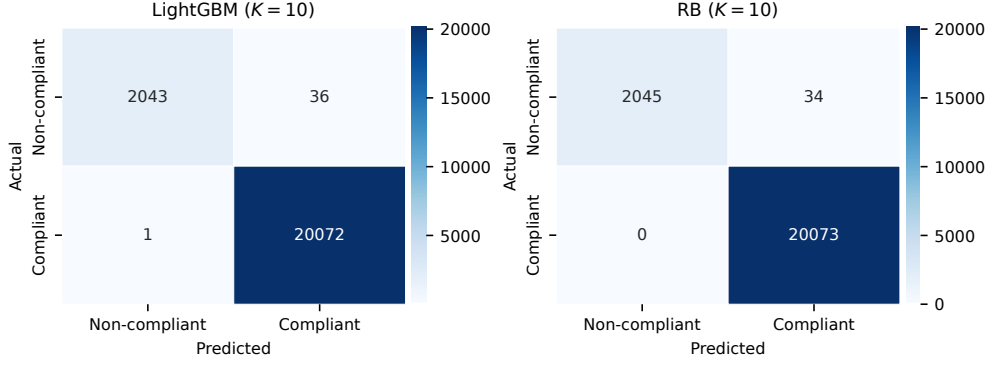


Fig. 6 Confusion matrices for LightGBM at $K = 10$ and the rule-based approach at $K = 10$ over the full dataset, showing correctly and incorrectly classified samples by compliance label.

the full dataset and the rule-based approach 34. The key result is the distribution of these errors between the approaches: both misclassify the same 34 samples, LightGBM misclassifies 3 additional samples, and the rule-based approach makes no error that LightGBM does not also make. This distribution is summarised in Table 4 and shown as confusion matrices in Figure 6.

McNemar’s test for the paired comparison of errors returns $p = 0.500$, so the difference in the number of errors is not statistically significant. The substantive finding is that both approaches err on practically the same samples. If the approaches were complementary in their errors, their combination (for example, in the form of the two-step model) could be substantially better than either component; the data show the opposite, as the overlap of errors is almost complete, which directly limits any improvement achievable by a combined approach. This finding is central to interpreting the two-step model results in Section 4.7.

Analysis at individual values of K reveals a further side result. At low K , specifically at $K = 10$, LightGBM attains $F_1 = 0.995 \pm 0.003$, the highest among the machine-learning models at this attribute-set size (Gradient Boosting reaches 0.994 ± 0.004 and Random Forest 0.993 ± 0.002). This advantage reflects two design features of LightGBM (Section 3.5.2), Gradient-based One-Side Sampling and leaf-wise tree growth: with a small attribute set, where the signal is already strongly concentrated, leaf-wise growth is more efficient. Nonetheless, even LightGBM does not surpass the rule-based approach, which attains $F_1 = 0.995 \pm 0.003$ at the same K ; the difference is within one thousandth of the F_1 measure and entirely insignificant under McNemar’s test. In practice, the rule-based approach remains the first choice on account of its interpretability and independence from the training process, while LightGBM is the most reasonable alternative when only a limited number of parameters is available and the rule-based approach cannot be used.

4.5 Effect of derived features

This short subsection examines the effect of the derived total pesticide concentration sum (Section 3.3), addressing RQ2. Adding the derived feature does not improve the F_1 measure for any approach, at any value of K , to a statistically significant extent. In the base scenario without the derived feature ($K_{\max} = 68$), Gradient Boosting attains $F_1 = 0.997$, and in the scenario with the added feature ($K_{\max} = 69$) it again attains $F_1 = 0.997$, the difference being within the standard deviation. The main reason is the rarity of the total pesticide sum measurements: this value is non-zero for only 486 of the 22,152 samples (2.2% of the dataset). At such limited coverage, the attribute cannot contribute to model training to a degree that would show in average performance. Moreover, its SHAP importance in the scenario with the derived feature is only 25th ($\bar{\phi} = 0.038$), below the inclusion threshold for all subsets $K \leq 20$. Interestingly, in the scenario with the metolachlor regulatory correction its importance rises to 12th

K	GB			RB		
	before	after	Δ	before	after	Δ
20	0.997	0.997	+0.000	0.997	0.997	+0.000
30	0.997	0.997	+0.000	0.998	0.998	+0.001
40	0.997	0.997	+0.000	0.996	0.999	+0.002
68	0.997	0.997	+0.000	0.997	0.998	+0.002

Table 5 F_1 measure before and after the metolachlor regulatory correction at selected values of K , averaged over the seven LOYO folds. The Δ column is the absolute difference (F_1 after minus F_1 before). Positive differences are in bold.

place ($\bar{\phi} = 0.093$), because removing the regulatory-invalid metolachlor measurements makes the total pesticide sum relatively more informative; even so, the attribute does not enable the machine-learning models to surpass the rule-based approach.

4.6 Effect of the metolachlor regulatory correction

This subsection presents the effect of the proposed metolachlor regulatory correction (Section 3.3) on the performance of all approaches and on their relationship, addressing RQ3 and representing the most important quantitative contribution of this work. Table 5 compares performance before and after the correction at the most informative values of K .

The correction affects the two main approaches asymmetrically. For the machine-learning models, the absolute change in F_1 remains within measurement precision and is practically undetectable. For the rule-based approach, the correction brings a noticeable improvement at $K \geq 30$, where the F_1 measure increases by 0.001 to 0.002 units. The absolute difference seems small, but in the context of a dataset with 22,152 samples and only 2,079 non-compliant ones it means the removal of several false positives. The best result of the whole experiment is the rule-based approach at $K = 40$ with the correction, which makes only 9 errors over the full dataset, all of them false negatives (0 false positives, 9 false negatives, aggregated over all seven LOYO folds). Compared with the best base-scenario result (the rule-based approach at $K = 30$ with 19 errors), the correction more than halves the number of errors. At

1243 this configuration, the average precision is 0.9998, the average recall 0.9977, and the
1244 F_1 measure 0.9988. On a dataset of this size, this result represents a limit below which
1245 further improvement is possible only by reducing the number of false negatives, which
1246 cannot be done without risking an increase in false positives.

1247
1248
1249 The statistical comparison between the rule-based approach and the machine-
1250 learning side is presented in Section 4.7, where the rule-based approach is tested
1251 against the two-step model in both scenarios.

1252
1253
1254 The sign of the confidence interval reveals that after the correction the difference
1255 favours the rule-based approach, not the machine-learning models. This is counter-
1256 intuitive at first sight: an intervention designed to improve data quality ultimately
1257 strengthens the simpler approach more than the more complex models. The reason lies
1258 in the nature of the inconsistency removed. Before the correction, samples from 2019–
1259 2024 with metolachlor ESA or OXA values above the $0.1\mu\text{g/l}$ threshold generated
1260 false positives for the rule-based approach, which applied the threshold without regard
1261 to the sampling year; after the correction these values are no longer in the dataset,
1262 which directly removes this entire group of rule-based errors. For the machine-learning
1263 models the influence of the same samples was smaller, as the models could in some
1264 cases override the predicted non-compliance through other attributes.

1273 1274 4.7 Two-step model

1275
1276 This subsection presents the evaluation of the two-step model (Section 3.5.3) in both
1277 main scenarios, addressing RQ4. The two-step model was evaluated in combination
1278 with all three tree-based machine-learning models (GB, RF, LightGBM); the most
1279 informative results are shown for the Gradient Boosting configuration, which achieves
1280 the highest performance in the base scenario (Table 6).

1281
1282
1283 In the base scenario, the two-step model attains $F_1 = 0.997$ at $K = 68$, the highest
1284 value among all approaches in that scenario, but the difference from the rule-based
1285 approach is small.

K	Without correction			With correction		
	GB	2step-GB	RB	GB	2step-GB	RB
20	0.997	0.997	0.997	0.997	0.997	0.997
40	0.997	0.997	0.996	0.997	0.997	0.999
68	0.997	0.997	0.997	0.997	0.997	0.998

Table 6 Comparison of the two-step model (2step-GB), the standalone Gradient Boosting, and the rule-based approach at selected values of K , showing the F_1 measure in the base scenario and in the scenario with the metolachlor regulatory correction.

Scenario	K	McNemar p	95% CI ($F_1^{2step} - F_1^{RB}$)	t -test p	Label
Without correction	68	0.324	$[-0.0007, +0.0025]$	0.401	n.s.
With correction	40	0.002	$[-0.0022, -0.0005]$	0.023	sig.

Table 7 Statistical tests of the difference in F_1 between the two-step model (2step-GB) and the rule-based approach, in both scenarios, each at the attribute-set size at which the respective scenario is most informative. The McNemar p column refers to McNemar’s test at the level of individual predictions, and the t -test p column to the paired t -test on the seven per-fold F_1 values. “sig.” denotes statistically significant and “n.s.” not significant at $\alpha = 0.05$.

approach is not statistically significant (McNemar’s test $p = 0.324$, 95% confidence interval $[-0.0007, +0.0025]$). In the scenario with the regulatory correction the picture changes: the rule-based approach becomes statistically significantly better than the two-step model at $K = 40$ (McNemar’s test $p = 0.002$, 95% confidence interval $[-0.0022, -0.0005]$). Both comparisons are summarised in Table 7. In both cases the two-step model therefore does not surpass the rule-based approach, which is a direct answer to RQ4.

The reason follows directly from the finding in Section 4.4: the errors of the two approaches largely overlap. The two-step model exploits the complementarity of the approaches only when they are independent in their errors; as both approaches err on the same samples here, the two-step model has no signal to exploit. In practice, when the two approaches agree the combined prediction does not differ from that of an individual approach, whereas when they disagree the two-step model risks adding further false negatives.

1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380

Approach	Precision	Recall	F_1
GB	0.9990	0.9954	0.9972
2step-GB	0.9995	0.9954	0.9975
RB	0.9998	0.9977	0.9988

Table 8 Precision and recall of the two-step model compared with the standalone Gradient Boosting and the rule-based approach at $K = 40$ in the scenario with the metolachlor regulatory correction.

Year	2019	2020	2021	2022	2023	2024	2025
ΔF_1	+0.002	+0.002	+0.001	+0.001	+0.003	0.000	0.000

Table 9 Difference $\Delta F_1 = F_1^{\text{RB}} - F_1^{\text{2step-GB}}$ across the individual LOYO folds at $K = 40$ in the scenario with the metolachlor regulatory correction. Positive values indicate better performance of the rule-based approach.

By design, the two-step model reduces the number of false positives, which is reflected in higher precision. Table 8 shows the precision and recall of the two-step model compared with the standalone approaches.

The two-step model indeed attains the highest precision among the evaluated machine-learning models (0.9995), but at the cost of lower recall compared with the standalone Gradient Boosting. The rule-based approach attains both the highest precision and the highest recall, a direct consequence of the correction removing the cause of false positives without affecting the ability to detect genuine non-compliances. This confirms the answer to RQ4: the complementarity that the two-step model could exploit is not present in the problem considered.

For a more detailed view of the temporal stability of the difference between the two-step model and the rule-based approach, Table 9 shows the difference in F_1 between the two approaches across the individual folds at $K = 40$ after the correction.

The difference ΔF_1 is positive or zero in all seven folds, meaning that the rule-based approach is not worse than the two-step model in any year. In 2024 and 2025 the

difference is zero, as expected: in 2025 all metolachlor measurements are regulatory-valid and the correction does not affect them, while in 2024 a relatively small share of samples has questionable values. The difference is largest in 2023 (+0.003), where the rule-based approach makes no error at this attribute set ($F_1 = 1.000$), and is also elevated in 2019 and 2020 (+0.002), consistent with the correction having the greatest effect where the data inconsistency is most pronounced.

4.8 Discussion

The significance of these results is twofold. On the one hand, they confirm that in routine drinking-water quality monitoring, where the domain is strongly regulatory-defined and the signal is concentrated in a few key indicators, the value of machine learning is limited. On the other hand, they show that domain-informed preprocessing interventions that address data inconsistencies can have a substantially greater effect on final performance than the choice of modelling approach. The negative outcome of the comparison between machine learning and the rule-based approach therefore does not reflect an inadequate methodology, but rather the nature of the dataset, in which the signal is already almost entirely captured by domain-defined thresholds.

The remaining share of errors in the best result is closely tied to labelling quality. A detailed inspection of the 9 errors of the rule-based approach at $K = 40$ with the correction gives the following breakdown: four are 2025 samples in which the metolachlor ESA value is measurably above the regulatory threshold of $0.1 \mu\text{g/l}$ yet the sample was nonetheless labelled compliant; three are other 2025 inconsistencies, two are from 2024, and one from 2021. For the samples with metolachlor ESA above the threshold, the label originates from the source record, so it could not be determined whether these are manual-entry errors, later labelling revisions, or an inconsistency between the laboratory measurement and the regulatory assessment. A methodologically important conclusion follows: at the level of performance reached (9 errors over

22,152 samples), the limiting factor for further progress is data quality rather than the choice of model. No choice of hyperparameters, no additional attribute, and no architecturally richer model can exceed the bound set by the labelling inconsistencies if these themselves represent the upper limit of what is achievable.

A further self-critical observation is that the rule-based approach and the best machine-learning model err on partly different samples in the corrected scenario at $K = 40$, where the near-complete overlap observed at $K = 10$ in the base scenario (Section 4.4) no longer holds. This represents a practical Pareto frontier between precision and recall: a further reduction in false positives is possible only at the cost of more false negatives, and vice versa. For drinking water, false negatives, that is, overlooked genuine non-compliances, are potentially dangerous from a public-health perspective, whereas false positives lead only to the additional cost of repeat sampling. In this sense, the rule-based approach in the scenario with the regulatory correction, which has 0 false positives and 9 false negatives, is not necessarily the optimal choice for practical use; a machine-learning model with an appropriately set decision threshold could move closer to the high-recall side by deliberately accepting a higher rate of false positives.

5 Conclusions

This work addressed the prediction of drinking-water sample compliance with the prescribed quality indicators, formulated as a binary classification problem on a dataset of 22,152 samples from 2019–2025. A rule-based reference approach and four machine-learning models (Gradient Boosting, Random Forest, LightGBM, and a multi-layer perceptron) were evaluated, with attribute selection performed by SHAP and performance assessed by LOYO time-series cross-validation. In addition to these established approaches, two contributions were proposed: the metolachlor regulatory correction

Scenario	Best approach	F_1	Errors
Base	RB ($K = 30$)	0.998	19
Derived features	GB ($K = 69$)	0.997	19
Regulatory correction	RB ($K = 40$)	0.999	9

Table 10 Summary of the best results across scenarios. Values are the F_1 measure and the number of errors over the full dataset (22,152 samples). The best values are in bold.

as a preprocessing step, and a two-step model combining the rule-based approach and a machine-learning model through a consensus requirement.

The best result of the whole experimental work is the rule-based approach in the scenario with the metolachlor regulatory correction at $K = 40$, which makes 9 errors over the full dataset ($F_1 = 0.999$), all in the form of false negatives. A summary of the best results across scenarios is given in Table 10. The proposed metolachlor regulatory correction thus more than halves the number of errors of the rule-based approach (from 19 to 9), and in none of the tested scenarios did the machine-learning models achieve a result that the rule-based approach did not achieve on its own. By design, the two-step model attains the highest precision among the machine-learning models (0.9995), but does not surpass the rule-based approach, as there is insufficient complementarity between the approaches and no additional discriminating signal is present in the data.

Read as a methodological result rather than a national case study, the finding is that a faithfully reconstructed rule-based baseline is an indispensable control when learned models are proposed for regulatory compliance tasks. Without it, the F_1 scores above 0.99 reported here for the machine-learning models would appear to be a success, when in fact they reproduce what the existing thresholds already achieve. The same applies to effort allocation: the single intervention that measurably improved accuracy was a domain-driven correction of a regulatory labelling inconsistency, not a change of model family or a hyperparameter search.

1519 Future work can proceed in several directions. The first is extending the dataset
1520 with additional sources, such as measurements from local water-supply utilities or
1521 environmental data on catchment areas, which could give the machine-learning models
1522 access to information not present in the current dataset and not covered by the pre-
1523 scribed indicators. The second is a systematic review of the labels of disputed samples
1524 by domain experts, which would directly address the identified labelling inconsisten-
1525 cies and thereby raise the upper bound on achievable performance. The third is the
1526 investigation of alternative cross-validation schemes, for example separated by region
1527 or water-source type, which would reveal any regional specificity of non-compliance
1528 patterns not detectable under the temporal LOYO scheme. A particularly interesting
1529 direction is the calibration of the decision threshold of the machine-learning models,
1530 which could deliberately select a point on the Pareto frontier with higher recall, a more
1531 appropriate choice than the current precision-oriented implementation in the context
1532 of protecting public health.
1533

1543 **Declarations**

1546 **Funding**

1549 This work was supported by the Water Research Center at the University of Ljubljana
1550 and by the University of Ljubljana Development Fund (RSUL). The funders had no
1551 role in the study design, in the collection, analysis or interpretation of the data, in the
1552 writing of the report, or in the decision to submit the article for publication.
1553

1556 **Competing interests**

1559 The authors declare that they have no competing interests as defined by Springer,
1560 or other interests that might be perceived to influence the results and/or discussion
1561 reported in this paper.
1562

Ethics approval and consent to participate	1565
	1566
Not applicable.	1567
	1568
	1569
Consent for publication	1570
	1571
Not applicable.	1572
	1573
	1574
Data availability	1575
	1576
The raw data are the public annual results of the Slovenian national drinking-water	1577
monitoring programme, published as seven Excel workbooks (2019–2025) at https://	1578
mpv.si/rezultati (Ministrstvo za zdravje, 2025). They are redistributed unmodified,	1579
together with the derived analysis dataset <code>clean_data_no_incomplete.csv</code> (22,152	1580
samples $\times$ 82 columns, 8.2 MB) and the aggregated result tables behind every	1581
figure and table in this article, at https://github.com/Aljaz20/water-compliance-ml	1582
(archived at https://doi.org/10.5281/zenodo.22798661 (Sovič, Hribar Lee, & Bosnić,	1583
2026)).	1584
	1585
	1586
	1587
	1588
	1589
	1590
	1591
Materials availability	1592
	1593
Not applicable: this study is a secondary analysis of existing monitoring records and	1594
did not involve the collection or generation of physical research materials.	1595
	1596
	1597
	1598
Code availability	1599
	1600
The analysis code (<code>water-compliance-ml</code>) was developed by A. Sovič (implementa-	1601
tion), with B. Hribar Lee and Z. Bosnić. Written in Python 3.12, using Jupyter and	1602
the packages pinned in <code>requirements.txt</code> (<code>pandas</code> , <code>numpy</code> , <code>scikit-learn</code> , <code>lightgbm</code> ,	1603
<code>torch</code> , <code>shap</code> , <code>optuna</code>); any CPU-only workstation is sufficient, no GPU is required.	1604
	1605
Free of charge under the MIT licence at https://github.com/Aljaz20/water-compliance	1606
	1607
	1608
	1609
	1610

1611 -ml, with the version underlying this article archived at [https://doi.org/10.5281/](https://doi.org/10.5281/zenodo.22798661)
1612 [zenodo.22798661](https://doi.org/10.5281/zenodo.22798661) (Sovič et al., 2026).
1613

1614

1615 **Author contributions**

1617 Conceptualization: A. Sovič, B. Hribar Lee, Z. Bosnić; Methodology: A. Sovič, Z.
1618 Bosnić; Software: A. Sovič; Formal analysis: A. Sovič; Investigation: A. Sovič; Data
1619 curation: A. Sovič; Validation: B. Hribar Lee; Supervision: B. Hribar Lee, Z. Bosnić;
1620 Writing – original draft: A. Sovič; Writing – review and editing: B. Hribar Lee, Z.
1621 Bosnić. All authors read and approved the final manuscript.
1622
1623
1624
1625
1626

1627 **Generative AI and AI-assisted technologies**

1628
1629 During the preparation of this work, the authors used a generative AI language tool
1630 in order to assist with language editing and to improve the readability and flow of
1631 the manuscript. After using this tool, the authors reviewed and edited the content as
1632 needed and take full responsibility for the content of the publication.
1633
1634
1635
1636

1637 **References**

- 1640 Ahmed, U., Mumtaz, R., Anwar, H., Shah, A., Irfan, R., García-Nieto, J. (2019,
1641 October). Efficient water quality prediction using supervised machine learning.
1642 *Water*, 11, 2210, <https://doi.org/10.3390/w11112210>
1643
1644
1645
1646
1647 Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019). Optuna: A
1648 next-generation hyperparameter optimization framework. *Proceedings of the*
1649 *25th ACM SIGKDD international conference on knowledge discovery and data*
1650 *mining* (pp. 2623–2631). ACM.
1651
1652
1653
1654
1655
1656

- Barzegar, Y., Gorelova, I., Bellini, F., D'Ascenzo, F. (2023). Drinking water
quality assessment using a fuzzy inference system method: A case study of
Rome (Italy). *International Journal of Environmental Research and Public
Health*, 20(15), 6522, <https://doi.org/10.3390/ijerph20156522> Retrieved from
<https://www.mdpi.com/1660-4601/20/15/6522>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32,
- Chen, W., Xu, D., Pan, B., Zhao, Y., Song, Y. (2024). Machine learning-based water
quality classification assessment. *Water*, 16(20), 2951, <https://doi.org/10.3390/w16202951> Retrieved from <https://www.mdpi.com/2073-4441/16/20/2951>
- Costa, D., Bayissa, Y., Barbosa, K.V., Villas-Boas, M.D., Bawa, A., Junior, J.L.,
... Srinivasan, R. (2024, November). Water quality estimates using machine
learning techniques in an experimental watershed. *Journal of Hydroinformatics*,
26(11), 2798-2814, <https://doi.org/10.2166/hydro.2024.132> Retrieved from
<https://doi.org/10.2166/hydro.2024.132>
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of
Statistics*, 7(1), 1–26, <https://doi.org/10.1214/aos/1176344552>
- Farzana, S.Z., Paudyal, D.R., Chadalavada, S., Alam, M.J. (2023). Prediction of water
quality in reservoirs: A comparative assessment of machine learning and deep
learning approaches in the case of Toowoomba, Queensland, Australia. *Geo-
sciences*, 13(10), 293, <https://doi.org/10.3390/geosciences13100293> Retrieved

1703 from <https://www.mdpi.com/2076-3263/13/10/293>
 1704
 1705
 1706
 1707 Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine.
 1708
 1709 *Annals of Statistics*, 29(5), 1189–1232,
 1710
 1711
 1712
 1713 Hu, X.C., Dai, M., Sun, J.M., Sunderland, E.M. (2023, March). The utility of machine
 1714
 1715 learning models for predicting chemical contaminants in drinking water: Promise,
 1716
 1717 challenges, and opportunities. *Curr. Environ. Health Rep.*, 10(1), 45–60,
 1718
 1719
 1720 Im, Y., Song, G., Lee, J., Cho, M. (2022). Deep learning methods for predicting
 1721
 1722 tap-water quality time series in South Korea. *Water*, 14(22), 3766, [https://](https://doi.org/10.3390/w14223766)
 1723
 1724 doi.org/10.3390/w14223766 Retrieved from [https://www.mdpi.com/2073-](https://www.mdpi.com/2073-4441/14/22/3766)
 1725
 1726 [4441/14/22/3766](https://www.mdpi.com/2073-4441/14/22/3766)
 1727
 1728
 1729
 1730 Katimon, A., Shahid, S., Mohsenipour, M. (2018, December). Modeling water quality
 1731
 1732 and hydrological variables using ARIMA: a case study of Johor River, Malaysia.
 1733
 1734 *Sustainable Water Resources Management*, 4(4), 991–998, [https://doi.org/](https://doi.org/10.1007/s40899-017-0202-8)
 1735
 1736 [10.1007/s40899-017-0202-8](https://doi.org/10.1007/s40899-017-0202-8) Retrieved from [https://doi.org/10.1007/s40899-017-](https://doi.org/10.1007/s40899-017-0202-8)
 1737
 1738
 1739
 1740
 1741 Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017).
 1742
 1743 LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural*
 1744
 1745 *information processing systems* (Vol. 30, pp. 3146–3154). Curran Associates, Inc.
 1746
 1747 Retrieved 2025-01-01, from [https://proceedings.neurips.cc/paper_files/paper/](https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf)
 1748
[2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf)

Kingma, D.P., & Ba, J. (2015). Adam: A method for stochastic optimization. *3rd international conference on learning representations (ICLR)*. San Diego, CA, USA. Retrieved from <https://arxiv.org/abs/1412.6980>

Lundberg, S.M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc. Retrieved 2025-01-01, from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Maiolo, M., & Pantusa, D. (2021). Multivariate analysis of water quality data for drinking water supply systems. *Water*, 13(13), 1766, <https://doi.org/10.3390/w13131766> Retrieved from <https://www.mdpi.com/2073-4441/13/13/1766>

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157, <https://doi.org/10.1007/BF02295996>

Ministrstvo za zdravje (2004). *Pravilnik o pitni vodi*. Retrieved 2025-12-01, from <https://pisrs.si/pregledPredpisa?id=PRAV3713> (Uradni list RS, št. 19/04, 35/04, 26/06, 92/06, 25/09, 74/15, 51/17)

Ministrstvo za zdravje (2025). *[dataset] Rezultati monitoringa pitne vode*. Retrieved 2026-01-01, from <https://mpv.si/rezultati> (Accessed: January 2026)

1795 Nacionalni inštitut za javno zdravje (2025). *Opisi kemijskih parametrov, ki*
 1796 *jih določamo v pitni vodi.* Retrieved 2026-03-01, from [https://nijz.si/wp-](https://nijz.si/wp-content/uploads/2025/01/Opisi-kemijskih-parametrov-ki-jih-dolocamo-v-pitni-vodi-vezija-5.pdf)
 1797 [content/uploads/2025/01/Opisi-kemijskih-parametrov-ki-jih-dolocamo-v-pitni-](https://nijz.si/wp-content/uploads/2025/01/Opisi-kemijskih-parametrov-ki-jih-dolocamo-v-pitni-vodi-vezija-5.pdf)
 1798 [vodi-vezija-5.pdf](https://nijz.si/wp-content/uploads/2025/01/Opisi-kemijskih-parametrov-ki-jih-dolocamo-v-pitni-vodi-vezija-5.pdf) (Verzija 5)
 1799
 1800
 1801
 1802 Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986). Learning representations by
 1803 back-propagating errors. *Nature*, 323(6088), 533–536, [https://doi.org/10.1038/](https://doi.org/10.1038/323533a0)
 1804 [323533a0](https://doi.org/10.1038/323533a0)
 1805
 1806
 1807
 1808
 1809
 1810 Sovič, A., Hribar Lee, B., Bosnić, Z. (2026). *[dataset] water-compliance-ml: code*
 1811 *and data for benchmarking machine learning against a regulatory rule-based*
 1812 *baseline for drinking-water compliance classification.* Zenodo. Retrieved from
 1813 <https://doi.org/10.5281/zenodo.22798661>
 1814
 1815
 1816
 1817 Tadesse, K., & Dinka, M. (2023, June). Water quality class modeling using machine
 1818 learning algorithms at Roodeplaat Dam, South Africa. *Preprints*, , [https://](https://doi.org/10.20944/preprints202306.1016.v1)
 1819 doi.org/10.20944/preprints202306.1016.v1
 1820
 1821
 1822
 1823
 1824
 1825 Vlada Republike Slovenije (2023). *Uredba o pitni vodi.* Retrieved 2025-
 1826 12-01, from [https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina/2023-01-](https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina/2023-01-1848/uredba-o-pitni-vodi)
 1827 [1848/uredba-o-pitni-vodi](https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina/2023-01-1848/uredba-o-pitni-vodi) (Uradni list RS, št. 61/23)
 1828
 1829
 1830
 1831
 1832
 1833
 1834
 1835
 1836
 1837
 1838
 1839
 1840