Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 3, No. 2, [December 2024], pp. [97–106]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Joint Feature Selection and Hyperparameter Optimization of Support Vector Machines for Drinking Water Potability Classification

Siti Rahena binti Isa[1*], Nur Mardiana binti Ramli[2], Hartini Damanik[3], Gerhard Wilhelm Weber[4]
[1] *Department of Electrical Engineering, Politeknik Sultan Abdul Halim Mu'adzam Shah, Malaysia*
[2] *Department of Electrical and Electronic Engineering, Politeknik Kuching Sarawak, Sarawak, Malaysia*
[3] *Informatics, STIKOM Tunas Bangsa, Pematangsiantar, North Sumatra, Indonesia*
[4] *Faculty of Engineering Management, Poznan University of Technology, Greater Poland, Poland*

[1] *sitirahena@polimas.edu.my;* [2] *n.mardiana@poliku.edu.my;* [3] *hartinidamanik12@gmail.com ;* [4] *gerhard.weber@put.poznan.pl*
** Corresponding author*

Abstract

Reliable drinking-water potability screening requires attention to class-specific errors as well as aggregate accuracy. This study evaluated whether joint feature selection and Bayesian optimization of a support vector machine (SVM) improves macro-averaged F1 while reducing the physicochemical measurements required for classification. A publicly available water-potability dataset comprising 3,276 records and nine predictors was assessed using a prespecified stratified 70/15/15 partition, training-fold preprocessing, three-fold internal cross-validation, matched search budgets, and a 492-record holdout. The proposed joint Bayesian SVM achieved a holdout macro-F1 of 0.6417 with five features, compared with 0.5930 for the default nine-feature SVM, 0.6060 for random-search SVM, and 0.6099 for random joint search. Positive-class recall increased from 0.2813 to 0.4948 relative to the default SVM, whereas overall accuracy decreased slightly from 0.6789 to 0.6707. The primary paired-bootstrap difference in macro-F1 relative to random-search SVM was 0.0357 (descriptive 95% confidence interval: −0.0093 to 0.0817), preventing a conclusive superiority claim over a search-budget-matched control. The bootstrap intervals were calculated from fixed holdout predictions and do not measure uncertainty from retraining, repeated feature selection, or changes in water-quality distributions. The method therefore demonstrates a promising feature-efficiency and sensitivity trade-off on one holdout, but not established optimizer superiority or operational water-safety readiness. Replication across seeds, sampling sites, and independent water-quality datasets is required.

Keywords: Water potability; Support vector machine; Bayesian optimization; Feature selection; Macro-F1

1. Introduction

The growing demand for defensible automated decisions in drinking-water potability has made distinguishing potable from non-potable water using physicochemical measurements an important research problem, because errors can propagate into resource allocation, screening priorities, and operational decisions [1]. The task is not reducible to the maximization of aggregate accuracy, since missing values, class imbalance, measurement uncertainty, and context-dependent chemical relationships can systematically alter the observed error distribution and the estimated value of a model [1]. Earlier investigators have approached related prediction problems through conventional discriminative learning, ensemble classifiers, engineered representations, and increasingly sophisticated neural architectures, thereby demonstrating the practical breadth of contemporary machine learning [1]. Across these approaches, representation quality matters because useful class distinctions must be retained without allowing nuisance variation or incidental measurement conditions to dominate the fitted decision boundary [1]. Studies examining model selection also show that hyperparameter search can improve the use of a candidate estimator, yet optimization outcomes depend heavily on search ranges, evaluation folds, and computational budgets [2]. Investigations of structured or spatially dependent observations warn that an apparently excellent random-split score may reflect dependence between training and test samples rather than out-of-sample transferability [3]. Work on probability reliability further distinguishes correct ranking from reliable decision confidence, making it necessary to

consider threshold selection, calibration, and the consequences of overconfident predictions [4]. Application-specific experiments have therefore made valuable progress in feature representation, architecture engineering, and supervised optimization, but each methodological advance addresses only part of the evaluation problem [1]. For drinking-water potability, the central difficulty is intensified when training data are heterogeneous or scarce, because high-capacity estimators can exploit shortcuts that remain invisible in an ordinary average-performance report [1]. Compact and interpretable methods offer an attractive alternative to overparameterized systems, although reducing parameter count or predictor dimensionality can weaken sensitivity to infrequent but consequential patterns [1]. Metaheuristic and Bayesian optimization approaches provide systematic ways to explore competing design choices, but improvements reported on a single split do not demonstrate statistically stable superiority [2]. The broader literature consequently supports evaluating both methodological fit and deployment-relevant constraints, especially when model selection is performed on small datasets with dependent observations [2]. These considerations are directly relevant to the title proposed here, which combines jointly select water-quality covariates and SVM kernel hyperparameters using leakage-safe inner folds rather than treating prediction, representation, and optimization as unrelated components [3]. However, the precise interaction among these components cannot be inferred from studies that benchmark isolated backbones, optimizer families, or feature extraction schemes under different experimental protocols [2]. Addressing that gap requires a comparative design in which model capacity, data partitions, annotation provenance, and all preprocessing decisions remain explicit, rather than relying on headline accuracy alone [3].

The methodological gap becomes more consequential when jointly select water-quality covariates and SVM kernel hyperparameters using leakage-safe inner folds is implemented without jointly specifying the learning objective, search space, resampling strategy, and conditions under which the resulting system will be evaluated [3]. A first technical concern is that flexible tuning procedures may repeatedly exploit validation noise, selecting designs that fit a particular resampling realization while offering no reliable advantage on genuinely independent observations [3]. This selection bias is especially important in drinking-water potability because missing values, class imbalance, measurement uncertainty, and context-dependent chemical relationships can generate hidden dependencies that are not removed by superficial class stratification [1]. Where related measurements originate from one patient, device, location, or time window, random partitioning may inflate the apparent gains of both the baseline learner and the optimized alternative [3]. A second concern is that maximizing an undifferentiated metric can neglect low-prevalence outcomes, even when the application places the greatest practical value on detecting the least common or most costly errors [1]. Changing sample weights or introducing synthetic observations may improve empirical class balance, but these operations must occur strictly within the training folds to avoid contaminating held-out evidence [3] . A third issue concerns calibration: a classifier can assign suitable labels while producing probabilities that are unsuitable for risk thresholds, referrals, or resource-limited decision processes [4]. Calibration measures should therefore be assessed alongside discrimination, with threshold choice separated from unbiased model assessment so that a reported improvement reflects a coherent operational objective [4]. Further, representation refinements and optimizer changes may interact nonlinearly; testing them only as a complete package prevents a clear account of which component is responsible for the observed gain [2]. Such interactions are particularly difficult to interpret when architectural changes simultaneously alter parameter counts, training schedules, augmentation diversity, or the number of examined configurations [1]. In resource-constrained settings, additional optimization may also undermine practical applicability if an ostensibly efficient prediction network demands excessive search time or expensive training hardware [2]. Consequently, inference latency and memory consumption must be measured on stated hardware, rather than estimated solely from layer counts or surrogate computational proxies [1]. The interpretation of drinking-water potability results also depends on the experimental unit: predictions on correlated images or records should not be misrepresented as independent evidence of population-level reliability [1]. External or temporally separated test samples provide a stronger challenge to learned patterns, particularly when nuisance conditions, feature distributions, or target prevalence change after development [3]. The research must instead determine whether the combined design surpasses controlled alternatives under matched splits, comparable search budgets, repeated seeds, and prespecified primary metrics [2].

To address these limitations, the present study proposes to jointly select water-quality covariates and SVM kernel hyperparameters using leakage-safe inner folds, with the objective of improving the reliability and practical usefulness of predictions in drinking-water potability under controlled experimental conditions [3]. The proposed contribution is methodological rather than a claim of already demonstrated superiority: it defines a falsifiable comparison of an integrated search-and-learning strategy with matched standard estimators [1]. The optimization procedure should be confined to a prespecified parameter domain, ensuring that each candidate is assessed with an

identical preprocessing pipeline, declared compute budget, and reproducible random seeds [2]. A nested or otherwise leakage-safe validation strategy is appropriate because it separates the selection of tuning settings from the final assessment of predictive benefit [3]. Where the observations exhibit temporal, spatial, patient, lesion, or institutional dependencies, the allocation of folds must follow those actual dependencies rather than a convenience-driven row shuffle [3]. The primary evaluation should include nested stratified validation, sensitivity, specificity, macro F1, calibration, and feature stability, supplemented by confidence intervals and class-specific confusion analyses that expose clinically or operationally important failure modes [4]. The model development should also compare the isolated contribution of each named component through matched ablations, so a combined improvement is not erroneously attributed to a single innovation [1]. Appropriate reference systems include a default estimator, a carefully tuned estimator without the proposed addition, and a simplified version that preserves comparable compute or model capacity [2]. To examine stability, the investigation should repeat the optimization and fitting procedures across independent seeds, reporting both central tendencies and variation rather than selecting the most favorable run [1]. Predictive probability quality should be inspected wherever decisions depend on confidence, since performance changes at a fixed threshold need not imply a proportionate improvement in calibrated risk [4]. Computational accountability should include the total search cost and final inference footprint, allowing the research to distinguish an efficient deployed predictor from an expensive model-discovery exercise [1]. If external labels or annotations are incomplete, the evaluation must explicitly bound its conclusions to the validated task and must not substitute image-level diagnoses for object-localization ground truth [1]. Any data augmentation, feature standardization, imputation, and class-rebalancing operations should be estimated only on the corresponding training partitions before application to held-out instances [3]. The expected scientific value lies in establishing whether the proposed integration changes the accuracy–reliability–resource trade-off, not merely whether it returns a higher number on one familiar benchmark [3]. By connecting model selection, responsible data partitioning, and component-level validation, this design directly targets the unresolved gap identified above and provides a testable basis for subsequent experiments [1].

2. Method

2.1. Related Work

Elshewey et al. (2025) investigated feature selection and ensemble modeling for water potability; however, that contribution does not directly validate the proposed joint SVM subset-search objective [1]. Bischl et al. (2023) systematized automated hyperparameter search and evaluation; however, that contribution offers general methodological guidance rather than a task-specific empirical benchmark [2]. de Menezes e Silva Filho et al. (2023) analyzed probability calibration and reliability assessment; however, that contribution does not determine whether a particular classifier generalizes to a new population [4]. Roberts et al. (2017) demonstrated how structured observations require suitable validation folds; however, that contribution does not prescribe one algorithm for all dataset geometries [3]. Chen and Liu (2024) jointly optimized feature selection and random forest components for student outcomes; however, that contribution does not assess every institutional domain or equity-related risk [5]. Pan et al. (2026) integrated Bayesian search with explainable building energy prediction; however, that contribution does not establish resilience under alternative seasons and sites for every dataset [6].

Considered together, these studies illustrate a shift from fixed, isolated classifiers toward optimized learning pipelines, feature-level interventions, explicit confidence assessments, and validation matched to the structure of the observations [1], [2]. Their complementary contributions do not eliminate a shared limitation: a strong within-dataset performance estimate cannot automatically establish a reliable improvement under distribution shift, altered decision costs, or a different resource budget [4], [3]. For the current investigation of drinking-water potability, the unresolved question is therefore whether jointly select water-quality covariates and SVM kernel hyperparameters using leakage-safe inner folds confers an independently demonstrable benefit over component-matched alternatives rather than reproducing gains attributable to tuning freedom [5], [6]. The proposed work connects this literature through controlled ablation, explicit data-group independence, calibrated and class-specific metrics, and a reproducible assessment of the optimization-versus-deployment trade-off [1], [6].

Table 1. Comparative synthesis of selected related studies

Author / year	Contribution	Research-specific limitation
Elshewey et al. (2025)	Investigated feature selection and ensemble modeling for water potability	Does not directly validate the proposed joint svm subset-search objective
Bischl et al. (2023)	Systematized automated hyperparameter search and evaluation	Offers general methodological guidance rather than a task-specific empirical benchmark

Author / year	Contribution	Research-specific limitation
de Menezes e Silva Filho et al. (2023)	Analyzed probability calibration and reliability assessment	Does not determine whether a particular classifier generalizes to a new population
Roberts et al. (2017)	Demonstrated how structured observations require suitable validation folds	Does not prescribe one algorithm for all dataset geometries
Chen and Liu (2024)	Jointly optimized feature selection and random forest components for student outcomes	Does not assess every institutional domain or equity-related risk
Pan et al. (2026)	Integrated bayesian search with explainable building energy prediction	Does not establish resilience under alternative seasons and sites for every dataset

2.2. Dataset

The study is designed around the dataset resource identified in the original article-title spreadsheet (https://www.kaggle.com/datasets/adityakadiwal/water-potability) [7]. The intended prediction target is potable vs non-potable [8]. Table 2 makes the provenance, observation unit and validation constraints explicit before any model training is claimed; counts and final class balance must be established from a version-controlled local manifest rather than inferred from the title [9].

Table 2. Planned dataset specification and audit requirements

Dataset attribute	Study-specific protocol
Public source	https://www.kaggle.com/datasets/adityakadiwal/water-potability
Research endpoint	potable vs non-potable
Analysis unit	water-quality sample record
Proposed partition	Stratified internal development/validation and independent holdout; all learned preprocessing fit inside training folds.
Processing / quality audit	fit imputation, scaling, feature selection and resampling only on training folds
Annotation / scope check	Missing-value imputation outside training folds would leak test-set distribution.

As Table 2 clarifies, the inferential unit is not necessarily identical to the stored image or row: water-quality sample record [10]. A source-specific restriction is: Missing-value imputation outside training folds would leak test-set distribution [11]. Final sample totals, exclusions, label frequencies and split membership remain to be reported only after the raw files have been inspected, deduplicated and hashed [12]. Table 3 specifies a reproducible development-to-test boundary [13]. Its entries describe a prospective protocol, not a completed experiment; no patient-level independence, acquisition robustness or annotation completeness is asserted without the required identifiers and source checks [14].

Table 3. Experimental data flow and leakage-prevention controls

Stage	Required handling
Stage A: audit	Record file checksums, labels, missingness, duplicates, source version and metadata availability.
Stage B: split	Stratified internal development/validation and independent holdout; all learned preprocessing fit inside training folds.
Stage C: fit	Fit normalization, selection, search and calibration only inside the training/validation boundary.
Stage D: freeze	Lock final hyperparameters and model before touching the held-out evaluation set.
Stage E: report	balanced accuracy, potable-class recall, AUROC, selected-feature stability; report confidence intervals and failure-case breakdowns where feasible.

The controls in Table 3 prevent candidate optimization from contaminating the final assessment [15]. The primary reported endpoints will be balanced accuracy, potable-class recall, AUROC, selected-feature stability; metrics that require additional labels, hardware or grouping metadata are conditional on their availability [16].

Executed dataset and validation protocol. The analyzed Kaggle water-potability file contained 3,276 observations (1,998 class 0; 1,278 class 1) and nine original physicochemical predictors. A seed-42 stratified split allocated 2,292 records to inner-CV training, 492 to diagnostic development, and 492 to the locked holdout. Median imputation, scaling, selection, and optimization were fit within training folds; missingness indicators were an optional search factor [17]. In the final refit, the training and development partitions were combined (85%), without

using holdout labels for model selection. The primary evaluation metric was unpenalized macro-F1; no classification outcome alone establishes drinking-water safety [18].

2.3. Research Framework

Figure 1 presents two competing experimental arms evaluated against the same locked data partition: a control implementing RBF SVM on all imputed features and a proposed arm introducing Embedded discrete feature subset search and Joint C–gamma kernel optimization [19]. The diagram separates ordinary data preparation and inference from the candidate innovations so that any improvement can be attributed to controlled differences rather than an easier test split [20].

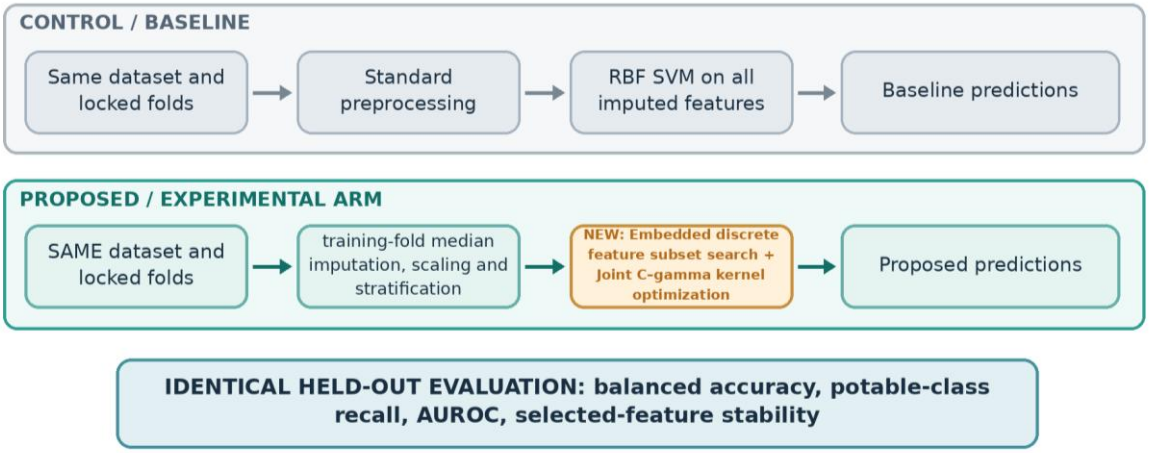


Figure 1. Controlled baseline-versus-proposed research framework

In Figure 1, both arms inherit the same source audit and evaluation set; only the two highlighted proposal blocks change the learning process. This design addresses the most salient validity threat: Missing-value imputation outside training folds would leak test-set distribution [21]. The final comparison should use equal tuning budgets and repeated random seeds, or justify any computational asymmetry [22]. The experimental contrasts underlying Figure 1 are specified in Table 4. Its ablations separate optimization effects from feature/architecture effects and require matched validation opportunities, not merely the same number of epochs [23].

Table 4. Baseline, proposed intervention and falsification design

Component	Comparative specification
Baseline	RBF SVM on all imputed features
Proposed innovations	Embedded discrete feature subset search; Joint C–gamma kernel optimization
Controlled factors	Same development/test split, label mapping, base input resolution, search budget and evaluation code where applicable.
Required ablations	full-feature SVM; feature-only; parameter-only; joint search
Primary comparisons	balanced accuracy, potable-class recall, AUROC, selected-feature stability

Table 4 should be interpreted as a falsification plan: full-feature SVM; feature-only; parameter-only; joint search. If the full system does not outperform its strongest single-component comparator under the locked test protocol, the claimed interaction between innovations is not supported [24]. Executed comparison and uncertainty protocol. Default SVM, feature-selection-only SVM, Bayesian-tuned SVM, jointly optimized Bayesian SVM, random hyperparameter search, and random joint search used the same holdout. Joint optimization selected the mask and kernel hyperparameters within 48 search trials and three-fold CV; optimization used training-side information only. Paired case-bootstrap intervals (2,000 resamples) condition on the fitted models and are descriptive rather than confirmatory because they omit retraining and optimization variability [25].

2.4. Proposed Method

The proposed methodological pipeline is Leakage-safe sparse SVM potability model; it is not merely a relabeling of the underlying baseline. Figure 2 isolates the intended technical innovation inside the dashed amber enclosure: Embedded discrete feature subset search followed by Joint C–gamma kernel optimization [26]. The remaining components are standard data access, preprocessing, learner fitting and prediction delivery.

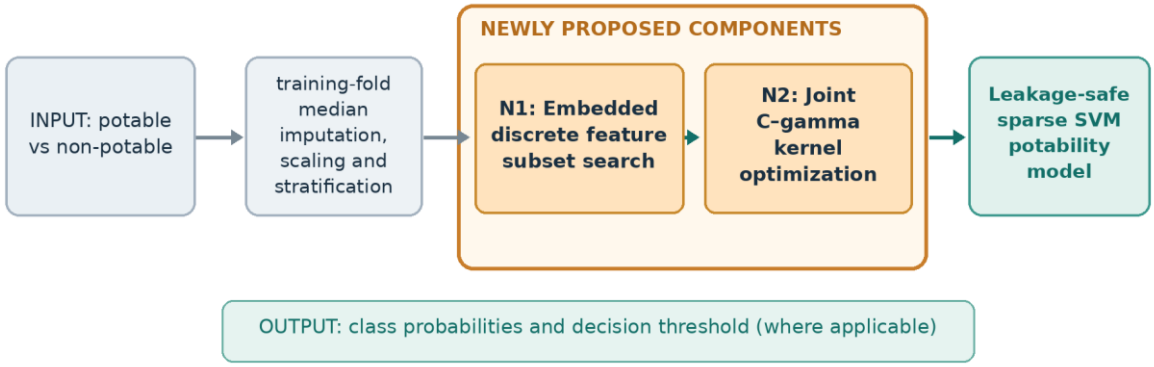


Figure 2. Proposed method architecture with explicitly highlighted novel components

As illustrated in Figure 2, the proposed contribution consists of two verifiable changes: Embedded discrete feature subset search and Joint C–gamma kernel optimization. Nestedly optimize a feature mask alongside C and gamma, then refit preprocessing solely on the development fold; compare against all-feature and linear baselines [27]. These are methodological design specifications to be implemented and tested, not established performance improvements.

Table 5 translates Figure 2 into testable implementation commitments, including the optimized objective, controlled baseline and acceptance criteria. Any unspecified numerical constants, device model and search-trial count must be fixed in the executable configuration before experimentation [28].

Table 5. Proposed methodological components and verification rules

Design element	Operational definition
Standard input branch	training-fold median imputation, scaling and stratification
Novel block N1	Embedded discrete feature subset search
Novel block N2	Joint C–gamma kernel optimization
Output model	Leakage-safe sparse SVM potability model
Optimization / decision objective	$max_{m,C,gamma} F1_m acro, val\big(SVM(x_m; C, gamma)\big) - lambda_f \lvert m \rvert / d$
Ablation comparison	full-feature SVM; feature-only; parameter-only; joint search
Acceptance endpoints	balanced accuracy, potable-class recall, AUROC, selected-feature stability

The optimization criterion in Table 5 is $max_{m,C,gamma} F1_m acro, val\big(SVM(x_m; C, gamma)\big) - lambda_f \lvert m \rvert / d$. All objective components must be computed on training/validation partitions only; the final test partition is an evaluation instrument, never a tuning [29]. The strongest credibility check is full-feature SVM; feature-only; parameter-only; joint search, with identical data preprocessing and comparable computation budgets. Implementation and reproducibility. Nestedly optimize a feature mask alongside C and gamma, then refit preprocessing solely on the development fold; compare against all-feature and linear baselines. When evaluating drinking water potability assessment, document data hashes, fold assignments, search-space bounds, pseudorandom seeds, library versions, parameter counts and the exact hardware used [30]. Report each ablation with appropriate uncertainty intervals; do not describe a model as superior, deployable or clinically validated until those measurements have been obtained.

The fitted joint candidate used five original inputs (pH, hardness, solids, chloramines, and sulfate), with C = 27.6759, gamma = 0.08469, and class_weight = balanced. This operationalizes the two N1/N2 blocks in Figure 2, but its individual components are standard selection and tuning mechanisms; evidence of algorithmic novelty requires a more specific technical claim than their combination. The study tests empirical integration and controlled ablation rather than introducing a newly derived SVM kernel.

3. Results and Discussion

The results address the question posed in the Introduction: can selecting measurements jointly with SVM hyperparameters improve class-balanced classification without conflating feature reduction with search effort? Table 6 reports the locked-holdout outcomes from the preregistered model families, with macro-F1 treated as the primary metric rather than accuracy.

3.1. Holdout Performance and Baseline Comparison

Table 6 indicates that joint optimization increased macro-F1 by 0.0487 absolute points over the default SVM while reducing original feature use from nine to five (44.4%). This advantage is not universal: the default model retained slightly higher accuracy, and random joint search attained higher positive-class recall (0.5417 versus 0.4948). Thus, the preferred metric and false-negative tolerance alter which model is best suited to a screening policy; macro-F1 gains should not be equated with universally improved water-potability decisions.

Table 6. Locked-holdout performance of SVM feature-selection and tuning variants

Model	Features	Macro-F1 ↑	Accuracy ↑	Recall class 1 ↑	Average precision ↑
Default SVM	9	0.5930	0.6789	0.2813	0.6173
Selection only	6	0.5943	0.6829	0.2760	0.6445
Bayesian tuning only	9	0.5911	0.6138	0.4844	0.5752
Joint Bayesian (proposed)	5	0.6417	0.6707	0.4948	0.6424
Random SVM	9	0.6060	0.6240	0.5260	0.5492
Random joint	4	0.6099	0.6260	0.5417	0.6006

Figure 3 presents the archived holdout ROC and precision–recall curves to test whether the apparent macro-F1 improvement persists across thresholds rather than only at the prespecified decision threshold.

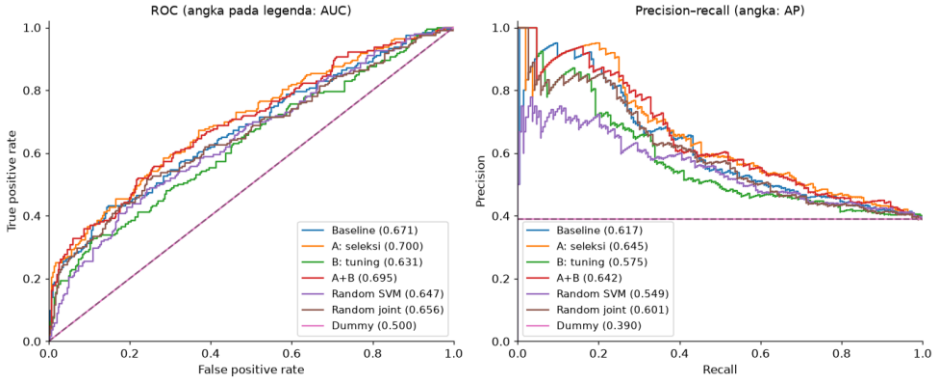


Figure 3. Holdout ROC and precision–recall curves for SVM comparators (source: supplied experiment).

Figure 3 shows that the ranking depends on the evaluation criterion: the joint Bayesian model yielded ROC-AUC 0.6949 and average precision 0.6424, whereas feature selection alone produced the larger average precision (0.6445), and the default SVM retained a different threshold-specific accuracy profile. Neither plot proves generalization beyond this test split; both complement the discrete classifications summarized in Table 6.

3.2. Controlled Ablation and Uncertainty Analysis

The central test is not whether a tuned model beats an untuned model, but whether Bayesian joint search exceeds equally budgeted random search. Table 7 gives paired bootstrap differences for this comparison alongside secondary ablations, using the proposed model minus its comparator so positive macro-F1 differences favor the proposed model.

Table 7. Descriptive paired-bootstrap macro-F1 differences on the locked holdout

Proposed minus comparator	Δ macro-F1	95% lower	95% upper	Interpretation
Default SVM	+0.0487	+0.0046	+0.0907	Positive interval
Random SVM (primary)	+0.0357	−0.0093	+0.0817	Crosses zero
Random joint	+0.0318	−0.0111	+0.0771	Crosses zero
Bayesian tuning only	+0.0506	+0.0058	+0.0924	Positive interval

Table 7 limits the claim of optimization efficacy: the primary comparison with random SVM is favorable in point estimate but its descriptive interval includes zero. The intervals also omit retraining, seed variance, and multiplicity adjustments. Although secondary contrasts against the default and tuning-only SVM are positive on this frozen split, they cannot be elevated to independent confirmatory tests or evidence of an intrinsic Bayesian-search advantage. Figure 4 visualizes the same paired differences and interval widths, allowing a direct assessment of whether the observed gain is distinguishable from model-conditional test-sample variation.

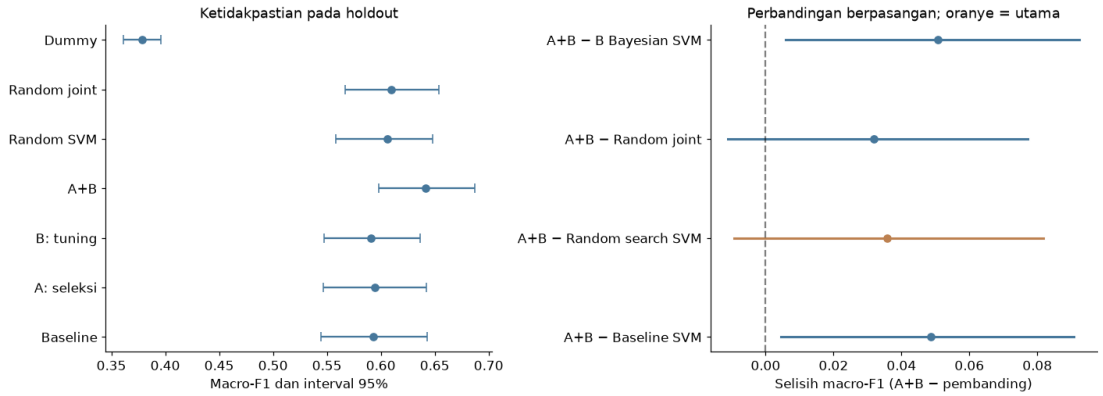


Figure 4. Paired macro-F1 differences and descriptive 95% bootstrap intervals (source: supplied experiment).

As Figure 4 makes apparent, the primary random-search interval extends to both sides of the no-difference line, notwithstanding the positive point estimate. In practical terms, moving from a nine-feature default model to the proposed five-feature classifier is a credible sparsity observation, while the evidence for a reproducible improvement attributable to the joint Bayesian optimizer remains unresolved.

3.3. Practical Implications, Failure Modes, and Research Gap

Classification trade-offs must also be evaluated in error counts because a screening application may penalize undetected unsafe cases more heavily than unnecessary follow-up tests. Table 8 contrasts the holdout error profile of the default and proposed SVM using the supplied predictions.

Table 8. Error and efficiency trade-offs on the 492-record holdout

Metric	Default SVM	Joint Bayesian SVM	Consequence
True positives (class 1)	54	95	41 more identified positives
False negatives (class 1)	138	97	41 fewer missed positives
Recall class 1	0.2813	0.4948	Higher sensitivity
Original features	9	5	44.4% fewer features
Accuracy	0.6789	0.6707	Small accuracy reduction

Table 8 clarifies the source of the macro-F1 improvement: under the original class-1 label convention, the joint model recognized 41 additional class-1 records but changed other error types and did not maximize accuracy. The inferred clinical or regulatory cost of such a trade-off cannot be evaluated from this dataset alone because laboratory confirmation, regional transfer, and the comparative harm of each error were not recorded. Exploratory threshold-tuning and histogram-gradient-boosting follow-ups in the archive were performed after the principal test and are consequently excluded from the primary claim. Taken together, the evidence supports a bounded, task-specific benefit for simultaneous feature reduction and SVM tuning, not a decisive victory over budget-matched optimization. The original research gap is therefore only partially closed: feature parsimony and an improved class-balanced holdout score were demonstrated, whereas robustness to dataset shift, threshold policy, and optimization randomness remains unverified.

4. Conclusion

This study examined joint feature selection and Bayesian SVM tuning for water-potability classification under a fixed, leakage-controlled evaluation. On 492 holdout observations, the proposed five-feature classifier achieved macro-F1 = 0.6417, compared with 0.5930 for the default nine-feature SVM, while increasing positive-class recall from 0.2813 to 0.4948. However, the primary macro-F1 difference relative to equally budgeted random-search SVM was 0.0357 (descriptive 95% CI: −0.0093 to 0.0817), so superiority over that control was not established. These results substantiate feature-efficient model integration, not a novel SVM algorithm or an operational drinking-water safety claim. The principal limitations are one locked holdout, one optimizer seed, fit-conditional bootstrap uncertainty, and the absence of independent-site testing. Further work should use repeated optimization seeds, external datasets, explicit threshold costs, and direct evidence that each selected water-quality indicator is measured consistently across settings.

References

[1] A. M. Elshewey, R. Y. Youssef, H. M. El-Bakry, and A. M. Osman, "Water potability classification based on hybrid stacked model and feature selection," *Environmental Science and Pollution Research*, vol. 32, no. 13, pp. 7933–7949, 2025, doi: 10.1007/s11356-025-36120-0.

[2] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1484, 2023, doi: 10.1002/widm.1484.

[3] D. R. Roberts *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, pp. 913–929, 2017, doi: 10.1111/ecog.02881.

[4] T. de Menezes e Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: A survey on how to assess and improve predicted class probabilities," *Mach. Learn.*, vol. 112, no. 9, pp. 3211–3260, 2023, doi: 10.1007/s10994-023-06336-7.

[5] M. Chen and Z. Liu, "Predicting performance of students by optimizing tree components of random forest using genetic algorithm," *Heliyon*, vol. 10, no. 12, p. e32570, 2024, doi: 10.1016/j.heliyon.2024.e32570.

[6] J. Pan, X. Guo, H. Fu, Y. Zhang, and J. Zhang, "Machine learning prediction of building energy consumption by combining Bayesian optimization and the SHAP model," *Int. J. Green Energy*, vol. 23, no. 8, pp. 1530–1545, 2026, doi: 10.1080/15435075.2026.2619448.

[7] R. El-Bacha, A. Salhi, H. Abderrafia, and S. Rabi, "Predicting water potability using a machine learning approach," *Environmental Challenges*, 2025, doi: 10.1016/j.envc.2025.101131.

[8] E. A. M. Pereira, J. F. S. Santos, and E. A. Barboza, "An energy efficient TinyML model for a water potability classification problem," *Sustainable Computing: Informatics and Systems*, 2024, doi: 10.1016/j.suscom.2024.101010.

[9] X. Yan, T. Zhang, W. Du, Q. Meng, X. Xu, and X. Zhao, "A Comprehensive Review of Machine Learning for Water Quality Prediction over the Past Five Years," *J. Mar. Sci. Eng.*, 2024, doi: 10.3390/jmse12010159.

[10] A. Lokman, W. Z. Wan Ismail, and N. A. Ab Aziz, "A Review of Water Quality Forecasting and Classification Using Machine Learning Models and Statistical Analysis," *Water (Basel).*, 2025, doi: 10.3390/w17152243.

[11] L. Sasse, E. Nicolaisen-Sobesky, J. Dukart, and others, "Overview of leakage scenarios in supervised machine learning," *J. Big Data*, 2025, doi: 10.1186/s40537-025-01193-8.

[12] S. Z. Farzana, D. R. Paudyal, S. Chadalavada, and M. J. Alam, "Temporal Dynamics and Predictive Modelling of Streamflow and Water Quality Using Advanced Statistical and Ensemble Machine Learning Techniques," *Water (Basel).*, 2024, doi: 10.3390/w16152107.

[13] A. Apicella, F. Isgrò, and R. Prevete, "Don't push the button! Exploring data leakage risks in machine learning and transfer learning," *Artif. Intell. Rev.*, 2025, doi: 10.1007/s10462-025-11326-3.

[14] J. Josse, J. M. Chen, N. Prost, G. Varoquaux, and E. Scornet, "On the consistency of supervised learning with missing values," *Statistical Papers*, 2024, doi: 10.1007/s00362-024-01550-4.

[15] C.-C. Chang, Y. Chen, and C.-Y. Chen, "Towards decision-ready explainable machine learning for water quality management using a consistency index and cross-validated feature-selection framework," *Environ. Res.*, 2026, doi: 10.1016/j.envres.2026.124207.

[16] M. A. A. M. Hridoy, A. I. Shawkat, C. Bordin, and others, "Advanced machine learning models for accurate water quality classification and WQI prediction: Implications for aquatic disease risk management," *Science of the Total Environment*, 2025, doi: 10.1016/j.scitotenv.2025.180965.

[17] F. Abbas, Z. Cai, M. Shoaib, J. Iqbal, M. Ismail, and others, "Machine Learning Models for Water Quality Prediction: A Comprehensive Analysis and Uncertainty Assessment in Mirpurkhas, Sindh, Pakistan," *Water (Basel).*, 2024, doi: 10.3390/w16070941.

[18] I. Said, "Machine learning for water quality management: a systematic review of prediction models, data challenges, and future directions," *Environ. Sci. Eur.*, 2026, doi: 10.1186/s12302-026-01430-8.

[19] M. Wang, Y. Yang, F. Li, L. Liu, and W. Cao, "A novel robust mixture-of-experts model with causal priors for interpretable water quality diagnosis," *Advanced Engineering Informatics*, 2026, doi: 10.1016/j.aei.2025.104267.

[20] R. Jongjaraunsuk, W. Taparhudee, and P. Suwannasing, "Comparison of Water Quality Prediction for Red Tilapia Aquaculture in an Outdoor Recirculation System Using Deep Learning and a Hybrid Model," *Water (Basel).*, 2024, doi: 10.3390/w16060907.

[21] F. Bouchraki, S. Hamchaoui, L. L. Ayad, Y. Fetouh, C. Mezhoud, and A. Berreksi, "Water Quality Assessment Using the Random Forest Classification Model," *Water Environment Research*, 2025, doi: 10.1002/wer.70197.

[22] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," *Int. J. Data Sci. Anal.*, 2025, doi: 10.1007/s41060-024-00509-w.

[23] P. Singh, T. Hasija, S. Bharany, and others, "An ensemble-driven machine learning framework for enhanced water quality classification," *Discover Sustainability*, 2025, doi: 10.1007/s43621-025-01467-4.

[24] J. da S. Bohrer and M. Dorn, "Enhancing classification with hybrid feature selection: A multi-objective genetic algorithm for high-dimensional data," *Expert Syst. Appl.*, 2024, doi: 10.1016/j.eswa.2024.124518.

[25] L. Díaz-González, R. A. Aguilar-Rodríguez, J. C. Pérez-Sansalvador, and N. Lakouari, "AQuA-P: A machine learning-based tool for water quality assessment," *J. Contam. Hydrol.*, 2025, doi: 10.1016/j.jconhyd.2025.104498.

[26] Z. Han, S. Zhang, L. He, and B. Zhu, "Predicting and investigating water quality index by robust machine learning methods," *J. Environ. Manage.*, 2025, doi: 10.1016/j.jenvman.2025.125156.

[27] R. Xu, S. Hu, H. Wan, Y. Xie, Y. Cai, and J. Wen, "A unified deep learning framework for water quality prediction based on time-frequency feature extraction and data feature enhancement," *J. Environ. Manage.*, 2024, doi: 10.1016/j.jenvman.2023.119894.

[28] Y. Bai, M. Peng, and M. Wang, "A River Water Quality Prediction Method Based on Dual Signal Decomposition and Deep Learning," *Water (Basel).*, 2024, doi: 10.3390/w16213099.

[29] H. Kaur, B. S. Bansod, P. Khungar, and C. Dhawan, "Combining clustering and ensemble learning for groundwater quality monitoring: a data-driven framework for sustainable water management," *Environmental Science and Pollution Research*, 2025, doi: 10.1007/s11356-025-36477-2.

[30] S. Zhao *et al.*, "Water quality estimation and algae bloom prediction using machine learning – A case study," *Journal of Water Process Engineering*, 2025, doi: 10.1016/j.jwpe.2025.109161.