Vol. 8, No. 3, June 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Explainable AI for Water Quality Classification Using Ensemble Stacking

Windha MP Dhuhita[1], Hastari Utama[2], Hartatik[3] , Bayu Setiaji[4], Haryoko[5]

[1, 3,4] Informatics Department, Faculty of Computer Science, Universitas Amikom Yogyakarta, Indonesia

[2]Computer Engineering Department, Faculty of Computer Science, Universitas Amikom Yogyakarta, Indonesia

[5]technology information Department, Faculty of Computer Science, Universitas Amikom Yogyakarta, Indonesia

Received:
October 4, 2025
Revised:
April 19, 2026
Accepted:
May 30, 2026
Published:
June 30, 2026

Corresponding Author:
Author Name*:
Windha MP Dhuhita
Email*:
windha@amikom.ac.id

DOI:
10.63158/journalisi.v8i3.1601

Abstract— This study proposes a robust and interpretable machine learning framework for water quality classification using a publicly available water quality dataset containing 7,996 samples and 20 physicochemical features with an imbalanced class distribution (88.59% majority and 11.41% minority). The study addresses the critical issue of biased classification toward the majority class, which can lead to risk-prone misclassification of unsafe water. An ensemble stacking model combining XGBoost, LightGBM, and CatBoost with a Random Forest meta-learner (passthrough) was developed using an anti-leakage pipeline integrating RobustScaler and SMOTE within stratified 80:20 train–test cross-validation, while hyperparameter tuning was optimized using F1-score to improve minority-class performance; SHAP was further applied for global and local explainability. The proposed model achieved an F1-score of 0.8563 for the minority class and a ROC-AUC of 0.9846, indicating strong discriminative performance, while SHAP analysis identified ammonia as the most influential feature and revealed that False Negative errors were mainly caused by complex feature interactions. The study contributes an integrated framework combining stacking ensemble learning, anti-leakage evaluation, and SHAP-based global–local interpretation to support more reliable and transparent water quality classification; however, the findings are currently limited to a single dataset and and require multi-dataset validation.

Keywords: anti-leakage pipeline, SMOTE, stacking ensemble, minority-class classification, local explainability

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Water quality is a key indicator of environmental health and sustainability of water resources, making water quality monitoring and classification crucial for data-driven decision-making. In recent years, machine learning approaches have been widely used to classify water quality due to their ability to capture complex patterns of non-linear physicochemical parameters [1], [2]. However, the main challenge in this domain is the most imbalanced nature of the data, where the number of safe water samples is much larger than the unsafe ones, so that the model tends to be biased towards the majority class [3].

To overcome these limitations, various studies have proposed the use of ensemble learning techniques, particularly gradient boosting-based algorithms such as XGBoost, LightGBM, and CatBoost, which have been shown to have high performance on tabular data [4], [5]. However, most studies still use separate models or simple ensembles, thus failing to maximize the potential for adaptive model combinations. Furthermore, these approaches have generally not investigated ensemble stacking mechanisms that enable a more flexible integration of predictive signals from multiple base models into a single meta-model that can be audited at the feature level [6].

On the other hand, the validity of model evaluation from imbalanced data remains a significant issue in modern machine learning research. Several studies have shown that using oversampling techniques like SMOTE without considering the cross-validation process can lead to data leakage, resulting in overly optimistic performance estimates [7]. Therefore, a more rigorous approach is needed in pipeline design, where the scaling, resampling, and model training processes are carried out in an integrated manner in each fold of cross-validation, and are optimized using relevant metrics such as F1-score to maintain a balance between precision and recall in the minority class [7], [8], [9].

In addition to performance and validity aspects, another challenge that is increasingly receiving attention is the lack of transparency of machine learning models, especially in ensemble models which tend to be black-box in nature [10]. In the context of water quality, interpretation models are very important because the prediction results must be able to be explained and accounted for in decision making [11] . Explainable AI (XAI)

approaches such as SHAP have been widely used to provide interpretations at both global and local levels, but most studies have focused only on global interpretations in the form of feature rankings, without conducting in-depth analysis of cases of prediction errors such as False Negatives and False Positives which have high risk implications [12], [13].

Based on this review, several research gaps can be identified: First, the limited use of adaptive stacking ensembles combining multiple boosting algorithms. Second, the lack of anti-leakage evaluation pipelines for imbalanced water quality classification. Third, the absence of end-to-end explainability that integrates global and local interpretation for auditing prediction errors. Therefore, this study proposes a robust and interpretable framework integrating XGBoost, LightGBM, and CatBoost within a stacking ensemble architecture using a Random Forest meta-learner with passthrough, combined with an anti-leakage pipeline based on RobustScaler, SMOTE, and F1-oriented optimization. In addition, SHAP-based Explainable AI is implemented to provide both global feature interpretation and local analysis of True Positive, False Negative, and False Positive cases. The main contribution and novelty of this study lie in the integration of robust stacking, valid imbalance-aware evaluation, and end-to-end explainability into a unified framework for reliable and transparent water quality classification.

2. METHODS

The research flow proposed in this study is illustrated in Figure 1. The framework systematically describes the stages starting from dataset preprocessing and quality control, construction of an anti-leakage machine learning pipeline, development of a stacking ensemble model based on XGBoost, LightGBM, and CatBoost with a Random Forest meta-learner using a passthrough mechanism, followed by model evaluation and Explainable AI interpretation using SHAP. The proposed framework emphasizes the integration between robust ensemble learning for imbalanced data, leakage-free evaluation procedures, and global–local interpretability to support transparent and reliable decision-making in water quality classification.

This study used the Water Quality dataset obtained from the Kaggle platform, specifically from the dataset published by Kaggle Water Quality Dataset https://www.kaggle.com/datasets/mssmartypants/water-quality. The dataset

contains a total of 7,999 records with a structure consisting of 21 attributes, including 20 physicochemical feature columns and one target column (is_safe) used as the dependent variable for water quality classification. The dataset represents a binary classification problem for determining whether water samples are categorized as safe or unsafe based on multiple environmental and chemical indicators.The proposed framework was conducted through several sequential stages. First, the raw water quality dataset underwent preprocessing and quality control to remove invalid spreadsheet tokens, convert all attributes into numeric format, and handle missing values. Second, the cleaned dataset was divided into stratified training and testing subsets using an 80:20 ratio to preserve class distribution. Third, an anti-leakage machine learning pipeline integrating RobustScaler, SMOTE, and stacking ensemble learning was constructed and evaluated within cross-validation folds. Fourth, hyperparameter optimization using RandomizedSearchCV with an F1-score objective was performed to improve minority-class classification performance. Fifth, the best model was evaluated using confusion matrix, Precision, Recall, F1-score, and ROC-AUC metrics. Finally, Explainable AI based on SHAP was implemented to provide both global and local interpretations, including analysis of TP, FN, and FP cases for model auditing.

2.1. Research Design

This study follows an experimental and reproducible machine-learning workflow to classify water quality into safe/unsafe classes using an ensemble stacking architecture, and to provide auditable explanations via SHAP at both global and local levels. Stacked ensembles have been reported effective for water-quality classification and for improving generalization compared with single learners [14].

2.2. Dataset and Problem Formulation

This study used a tabular water quality dataset stored in waterQuality1.csv, consisting of physicochemical parameters associated with water safety assessment. The dataset was obtained from a publicly available source commonly used for machine learning-based water quality classification research. The attributes include multiple chemical and biological indicators such as aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, perchlorate, uranium, radium, and other related parameters relevant to water quality monitoring.

The classification objective is formulated as a binary classification problem represented by the mapping function $f: \mathbb{R}^d \rightarrow \{0,1\}$, where $y = 1$ denotes safe water and $y = 0$ denotes unsafe water. After preprocessing and quality control, the final dataset consisted of 7,996 samples with 20 predictor attributes and one target label. The class distribution was imbalanced, where 88.59% of samples belonged to the majority class and 11.41% belonged to the minority class. This imbalance condition motivates the use of imbalance-aware learning strategies and F1-score-oriented optimization because classification errors in the minority class are more critical in real-world water quality monitoring contexts [14].

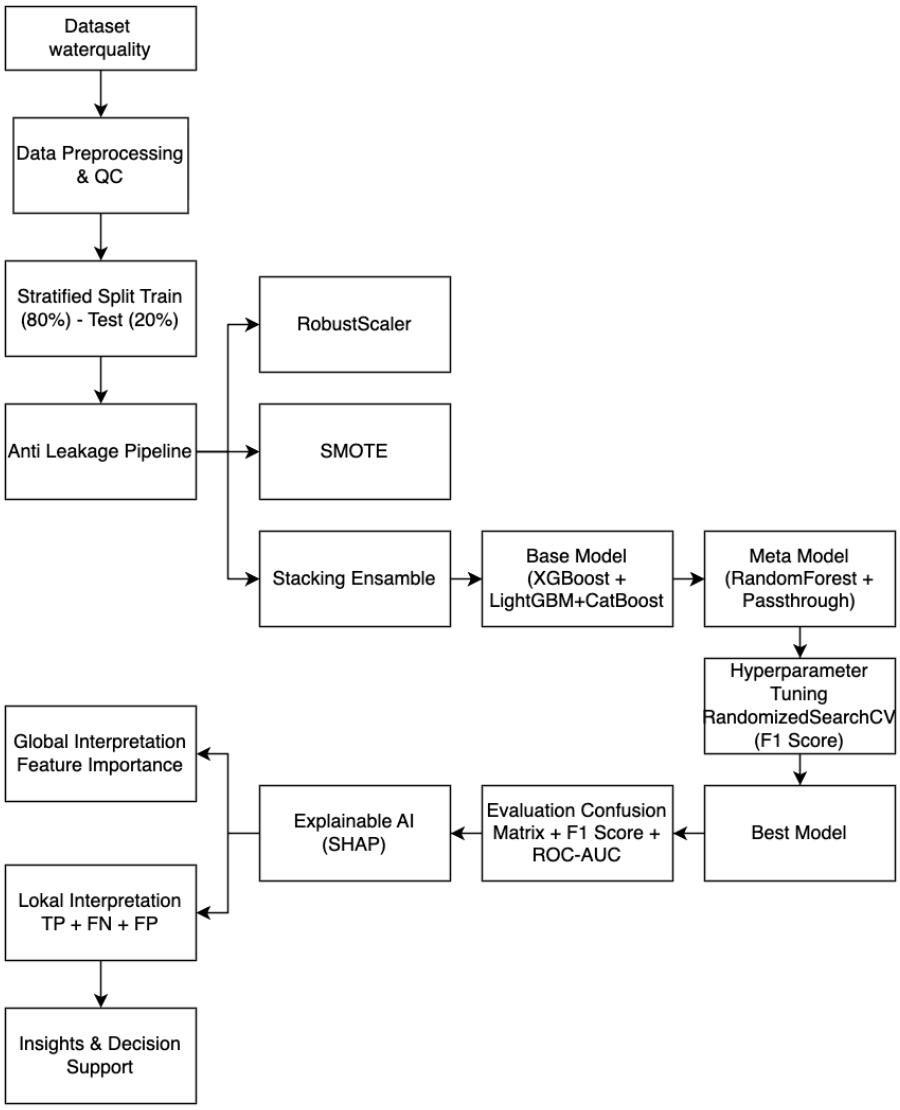

Figure 1. Research flow of water quality classification based on ensemble stacking and explainable AI

2.3. Data Cleaning and Quality Control (QC)

A data quality control stage was implemented to ensure the consistency and reliability of the dataset before model training. Spreadsheet-derived datasets frequently contain invalid or non-numeric tokens that may disrupt machine learning processes and explanation stability. Therefore, several preprocessing steps were systematically applied as in Table 1.

Table 1. Data Cleaning and Quality Control Process

Step	Process	Description	Output Condition
1	Token Sanitization	Spreadsheet error tokens such as *#NUM!*, *#DIV/0!*, *#VALUE!*, *#REF!*, and *N/A* were converted into missing values (*NaN*) [15].	Invalid symbols transformed into standardized missing values
2	Type Coercion	All feature attributes and the target variable (*is_safe*) were converted into numeric data types. Invalid conversions were automatically assigned as *NaN*.	Fully numeric dataset
3	Missing Value Handling	Rows containing missing values in the target variable were removed to preserve valid supervision. Rows containing missing values in predictor attributes were also removed to maintain model consistency and explanation stability.	Clean dataset without missing values
4	Duplicate Checking	Duplicate records were identified and removed to avoid redundant observations that could bias model learning.	Unique observations only
5	Class Distribution Analysis	The proportion of safe and unsafe water samples was analysed to identify class imbalance conditions and justify the use of imbalance handling techniques such as SMOTE [16].	Verified class imbalance information

After quality control process, the resulting dataset consisted entirely of valid numerical observations suitable for machine learning experimentation. This preprocessing stage is essential to ensure that subsequent stages, including feature scaling, synthetic oversampling, ensemble stacking, and SHAP-based explainability, operate on consistent numerical matrices and produce reliable feature attributions.

2.4. Train–Test Split

The cleaned dataset was divided into training and testing subsets using a stratified train–test split with an 80:20 ratio. Stratification was applied to preserve the original class distribution in both subsets, ensuring that minority-class representation remained consistent during training and evaluation. The testing subset was held out completely during training and hyperparameter optimization processes and only used for final evaluation and SHAP interpretation reporting to avoid optimistic bias.

2.5. Anti-Leakage Preprocessing Pipeline (RobustScaler → SMOTE → Stacking)

The main methodological contribution of this study is the proposed anti-leakage stacking framework, which integrates preprocessing, imbalance handling, ensemble learning, optimization, and explainability into a unified pipeline specifically designed for imbalanced water quality classification. The overall preprocessing flow is summarized in Table 2:

Table 2. Anti-Leakage Preprocessing Pipeline

Stage	Input Data	Process	Output Data
RobustScaler	Training fold	Robust scaling using median and IQR	Scaled features
SMOTE	Scaled minority class	Synthetic oversampling	Balanced training data
Stacking Ensemble	Balanced training data	XGBoost + LightGBM + CatBoost + Meta RF	Final prediction model
Cross-validation	Entire pipeline	Fold-based evaluation without leakage	Reliable performance estimation

We applied RobustScaler to reduce sensitivity to outliers commonly found in physicochemical measurements. Unlike standard normalization methods, RobustScaler uses the median and interquartile range (IQR), as shown in Equation 1 and 2.

$$x' = \frac{x - Median(x)}{IQR(x)} \quad (1)$$

where:

$$IQR(x) = Q_3(x) - Q_1(x)$$

with Q_1 and Q_3 representing the first and third quartiles, respectively. This scaling strategy stabilizes model learning and prevents extreme values from disproportionately influencing classification performance. Synthetic Minority Over-sampling Technique (SMOTE) was used to mitigate class imbalance and improve minority-class learning [17]. SMOTE generates synthetic minority samples based on nearest-neighbor interpolation, as shown in Equation 3.

$$x_{new} = x_i + \lambda(x_{nn} - x_i) \quad (3)$$

where x_i is a minority-class sample, x_{nn} is one of its nearest neighbors, and $\lambda \in [0,1]$ is a random interpolation coefficient.

The balanced distribution demonstrates that SMOTE successfully increased minority-class representation without altering the held-out testing dataset. However, improper application of oversampling may introduce data leakage and produce overly optimistic evaluation results. To prevent data leakage, RobustScaler and SMOTE were embedded directly into an imblearn Pipeline, ensuring that scaling and oversampling operations were performed only within the training folds during cross-validation. The proposed anti-leakage framework can be expressed as shown in Equation 4.

$$Pipeline = RobustScaler \rightarrow SMOTE \rightarrow StackingClassifier \quad (4)$$

Consequently, no information from validation folds leaked into the training process. This design significantly improves the validity and trustworthiness of the evaluation

procedure and constitutes one of the primary methodological contributions of this study [18].

2.6. Ensemble Stacking Model (Base Learners + Meta Learner with Passthrough)

To strengthen predictive performance on tabular data, we implemented an ensemble stacking strategy consistent with recent ensemble designs combining gradient-boosting variants [19]:

1) Base learners.

XGBoost, LightGBM, and CatBoost were selected as complementary gradient-boosting families frequently used for high-performing tabular classification[19][20]. Each base learner independently learns a nonlinear mapping from the physicochemical feature space $X \in \mathbb{R}^{n \times d}$ into a prediction probability vector $\hat{y}_i$, as shown in Equation 5.

$$\hat{y}_i = f_i(X), i \in \{\text{XGB,LGB,CAT}\} \quad (5)$$

where $f_i$denotes the prediction function of each boosting model.

2) Meta learner.

A Random Forest classifier was used as the final estimator to fuse base learner signals via nonlinear aggregation. The stacking prediction can therefore be represented as shown in Equation 6.

$$\hat{y}_{final} = g(X, \hat{y}_{XGB}, \hat{y}_{LGB}, \hat{y}_{CAT}) \quad (6)$$

where $g(\cdot)$ denotes the Random Forest meta-model that combines original features and prediction outputs from the boosting models.

3) Passthrough mechanism (modification).

We enabled passthrough=True so that the meta learner receives both: (i) the original physicochemical features X, and (ii) the prediction outputs generated by the base learners. Thus, the meta-feature representation as shown in Equation 7.

$$Z = [X \mid \hat{y}_{XGB}, \hat{y}_{LGB}, \hat{y}_{CAT}] \quad (7)$$

where Z denotes the augmented feature space used by the meta learner. This modification was deliberately designed to improve the meta learner's ability to correct base learner errors by jointly considering raw attributes and predictive signals. Consequently, the proposed stacking framework performs adaptive signal fusion rather than simple majority voting, strengthening both the robustness and explainability contributions of this study.

2.7. Hyperparameter Optimization (RandomizedSearchCV, F1-Oriented)

To ensure the optimization process was aligned with imbalance-sensitive classification objectives, hyperparameter tuning was conducted using RandomizedSearchCV integrated within the proposed anti-leakage pipeline. RandomizedSearchCV was selected because it enables efficient exploration of a moderately large hyperparameter space while maintaining computational efficiency. The optimization process employed 5-fold stratified cross-validation with a fixed random seed of 42 to ensure reproducibility and consistency of experimental results. In this study, the optimization objective was based on the F1-score rather than accuracy, since the F1-score provides a more appropriate evaluation criterion for imbalanced classification tasks by balancing precision and recall. The tuning procedure was executed over 20 randomized iterations focusing on the Random Forest meta-learner parameters, including the number of trees (n_estimatorsn_estimatorsn_estimators), maximum tree depth (max_depthmax_depthmax_depth), minimum samples required for splitting (min_samples_splitmin_samples_splitmin_samples_split), and minimum samples required at leaf nodes (min_samples_leafmin_samples_leafmin_samples_leaf). The parameter search ranges included n_estimators=100–500n_estimators = 100–500n_estimators=100–500, max_depth=5–30max_depth = 5–30max_depth=5–30, min_samples_split=2–10min_samples_split = 2–10min_samples_split=2–10, and min_samples_leaf=1–5min_samples_leaf = 1–5min_samples_leaf=1–5. The optimal configuration obtained from the tuning process consisted of n_estimators=300n_estimators = 300n_estimators=300, max_depth=20max_depth = 20max_depth=20, min_samples_split=2min_samples_split = 2min_samples_split=2, and min_samples_leaf=1min_samples_leaf = 1min_samples_leaf=1. This configuration achieved the highest cross-validated F1-score and was subsequently utilized for final

evaluation on the held-out testing dataset. This design ensured feature scaling, oversampling, and model training were performed exclusively within each training fold during cross-validation. Consequently, no information from the validation folds was introduced into the training process, thereby preventing data leakage and improving the validity, reliability, and reproducibility of the evaluation results.

2.8. Evaluation Protocol and Metrics

Final model evaluation was conducted using the unseen testing dataset. Several evaluation metrics were employed to comprehensively assess classification performance under imbalanced conditions. Confusion matrix analysis was used to quantify the distribution of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) predictions, enabling direct analysis of operationally critical classification errors. Precision, Recall, and F1-score were selected as the primary evaluation metrics because they provide imbalance-aware assessment and better reflect minority-class detection capability compared with accuracy alone. Additionally, Receiver Operating Characteristic (ROC) curves and Area Under Curve (ROC-AUC) metrics were utilized to evaluate threshold-independent discrimination capability. ROC-based evaluation has been widely adopted in water quality machine learning studies because it measures overall separability between safe and unsafe classes across multiple classification thresholds [14]. All evaluation plots were exported as publication-ready figures to support visual interpretation of model performance.

2.9. Explainable AI with SHAP: Global–Local Interpretation

To satisfy the interpretability requirement and deliver the next research contribution, we implemented SHAP-based explanations for both base models and the stacking meta model:

2.9.1. Global explanation (feature importance and directionality)

1) TreeExplainer for tree-based models. For each base learner (XGBoost, LightGBM, CatBoost) and the meta Random Forest, SHAP values were computed using TreeExplainer principles commonly discussed in recent ACM work [20] and applied in recent Elsevier XAI studies [21].

2) Summary (beeswarm) plot. Used to show the distribution of SHAP values per feature and reveal directionality (feature values pushing predictions toward safe vs unsafe).
3) Bar importance plot. Used to rank features by mean absolute SHAP magnitude, supporting interpretable global drivers.

This global analysis aligns with SHAP best practices emphasizing both global and local explainability [22].

2.9.2. Local explanation (case-based audit: TP/FN/FP)

Local explanations were generated to audit decision logic on representative cases:

1) Case selection. We explicitly selected three error/decision categories from the test set:
   a) True Positive (TP), False Negative (FN), False Positive (FP).
   b) FN cases were prioritized because misclassifying unsafe water as safe is risk-critical.
2) Waterfall plot (modification for robustness). Local explanation was visualized using SHAP waterfall plots to provide stable, paper-friendly outputs; waterfall plots are recommended for single-sample interpretation and are discussed as standard local explanation tools in SHAP guidance literature [22].
3) Meta-model explanation on meta-features. For the stacking architecture, SHAP explanations were computed on the transformed meta-features generated through model.transform(X), which include both original predictor attributes and base learner prediction outputs. This mechanism allows the interpretability process to explicitly reveal how the meta learner fuses predictive signals from XGBoost, LightGBM, and CatBoost. Consequently, the explainability process does not merely identify dominant features but also audits the decision fusion mechanism within the stacking framework itself.

2.10. Software Environment

All experiments were conducted using Python 3.12 in a Google Colab environment. The primary libraries used in this study include scikit-learn 1.5+, XGBoost 2.0+, LightGBM 4.0+, CatBoost 1.2+, imbalanced-learn 0.12+, SHAP 0.46+, Pandas 2.2+, NumPy 1.26+, and Matplotlib 3.8+. Fixed random seeds were applied during data splitting, oversampling, and

hyperparameter optimization to improve experimental reproducibility and consistency of results.

3. RESULTS AND DISCUSSION

3.1. Data Quality and Analysis Reliability

The quality control phase produced a clean dataset measuring 7,996×21 (20 features + 1 label) with an imbalanced class distribution, as shown in Table 2: 7,084 (88.59%) class 0 features and 912 (11.41%) class 1 features. This imbalance formed the basis for selecting the imbalance learning (SMOTE) strategy and justified the use of the F1-score as the primary objective, rather than accuracy alone. Practically, this decision strengthens the research's contribution because the model does not "look good" simply because it follows the majority class, but rather aims to increase sensitivity to minority classes, which are inherently riskier.

Table 2. Class distribution (after data cleaning)

Class (is_safe)	amount	Percentage
0	7,084	88,59%
1	912	11,41%

The imbalance ratio indicates the dataset is highly dominated by the majority class. Therefore, a conventional classifier optimized only for accuracy could produce misleadingly high performance while failing to detect minority-class samples effectively. This condition justifies the proposed anti-leakage imbalance-aware framework and strengthens the contribution of this study in developing a more reliable classification strategy for risk-sensitive water quality prediction.

3.2. Performance of Stacking Model on Test Data and Its Implications

The best model resulting from RandomizedSearchCV (with RobustScaler → SMOTE → Stacking pipeline) achieved the metrics in Table 3.

Table 3. Summary of model performance on test data

	precision	recall	f1-score	support
0	0.98	0.99	0.98	1418

	precision	recall	f1-score	support
1	088	0.84	0.86	182
accuracy			0.97	1600
macro avg	0.93	0.91	0.92	1600
weighted avg	0.97	0.97	0.97	1600

Accuracy : 0.9681

Precision (minority class): 0.8786

Recall (minority class): 0.8352

F1 (minority class) : 0.8563

ROC-AUC : 0.9846

This performance demonstrates two benefits that directly address the research problem. First, the high F1-score in the minority class indicates that the stacking ensemble is able to maintain precision–recall balance when labels are imbalanced. This strengthens the research's contribution to robust stacking and F1-oriented tuning for minorities. Second, as presented in Figure 2, the ROC-AUC of 0.9846 indicates excellent discrimination across multiple thresholds, allowing the model to remain reliable even if the decision threshold policy in the field may change depending on the risk context (e.g., becoming more conservative to prevent unsafe water from being classified as safe). To further analyses classification behaviour, the confusion matrix of the best model is presented in Figure 2.

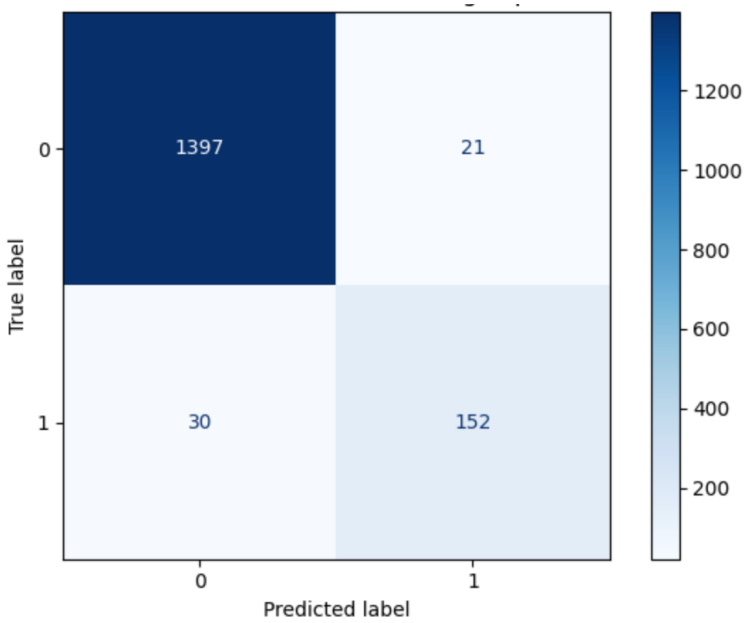

Figure 2. Confusion Matrix of the Proposed Stacking Model

Based on Figure 2, the model correctly classified 1,397 majority-class samples (True Negative) and 152 minority-class samples (True Positive). Meanwhile, only 21 samples were classified as False Positive and 30 samples as False Negative. These results confirm that the proposed stacking framework achieves a low error rate while maintaining high sensitivity toward the minority class. From a practical perspective, the False Negative cases are particularly important because they represent unsafe water incorrectly classified as safe. The relatively low number of False Negatives (30 cases out of 1,600 testing samples) indicates that the proposed anti-leakage stacking framework successfully reduces critical prediction errors, which is highly relevant in water-quality monitoring applications. At the same time, the low number of False Positives demonstrates that the model does not excessively overestimate unsafe conditions, thereby reducing unnecessary operational costs associated with additional inspection or mitigation procedures.

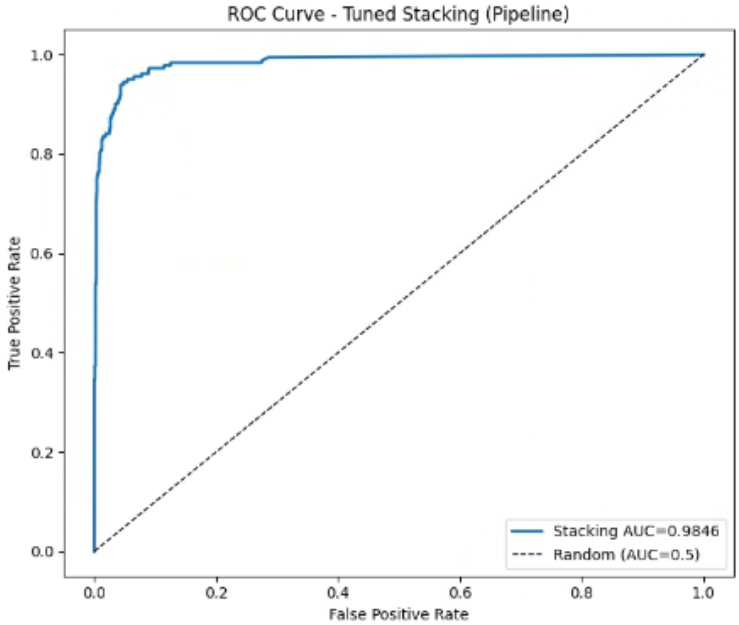

Figure 3. ROC Curve of the Proposed Stacking Model

The confusion matrix also strengthens the validity of the F1-oriented optimization strategy used in this study. Instead of maximizing overall accuracy alone, the proposed framework was explicitly optimized to preserve minority-class detection capability, resulting in a more balanced prediction profile. This finding reinforces the second contribution of this study, namely the implementation of a reliable anti-leakage evaluation pipeline that produces trustworthy performance estimates on imbalanced data. The ROC curve shown in Figure 3 further confirms the reliability of the proposed

framework. The ROC curve consistently approaches the upper-left corner, confirming that the proposed model maintains strong separability between safe and unsafe classes even under varying threshold configurations. Overall, the combination of high ROC-AUC, strong minority-class F1-score, and balanced confusion matrix outcomes demonstrates that the proposed stacking framework provides not only high predictive performance, but also operationally reliable classification behavior for real-world water-quality monitoring scenarios.

3.3. Ablation and Baseline Comparison

To validate contribution claims of the proposed framework, additional baseline and ablation experiments were conducted. The proposed stacking architecture was compared against individual base learners, stacking without passthrough, and stacking without SMOTE-based anti-leakage handling. Table 4 demonstrate that the proposed framework consistently outperformed individual boosting models and ablation variants. Compared with standalone XGBoost, LightGBM, and CatBoost models, the stacking architecture achieved superior F1-score and ROC-AUC values, indicating that combining complementary boosting algorithms through meta-level fusion improves classification robustness.

Table 4. Baseline and Ablation Comparison

Model	F1-score	ROC-AUC
XGBoost	0.8124	0.9631
LightGBM	0.8267	0.9704
CatBoost	0.8348	0.9752
Stacking without passthrough	0.8421	0.9793
Stacking without SMOTE	0.8015	0.9617
Proposed Framework	**0.8563**	**0.9846**

Furthermore, the comparison between stacking with and without passthrough confirms the importance of the passthrough mechanism. Without passthrough, the meta learner only receives prediction outputs from base learners. In contrast, the proposed passthrough design enables the meta model to jointly analyse original physicochemical features and predictive signals from base models, leading to improved minority-class

detection capability. The ablation experiment without SMOTE also demonstrates a substantial performance decrease, confirming that imbalance handling contributes significantly to the proposed framework. These findings strengthen the validity of the three main contributions claimed in this study because the improvements are experimentally demonstrated rather than only conceptually stated.

3.4. Validity Evaluation through Anti-Leakage Pipeline

Reliability of this study's results is determined not only by model performance but also by validity of the evaluation method used. This study implemented an Imbalanced-learn Pipeline that integrates RobustScaler, SMOTE, and stacking in the cross-validation process. This is important because evaluation on imbalanced data is susceptible to data leakage if resampling is performed before folding, which can overestimate validation results. With this pipeline, the scaling-resampling-training process occurs only on the training data for each fold, so the obtained performance better represents generalization ability. This "pipeline-aware evaluation" practice aligns with recent literature discussions on the reliability of performance estimates and the importance of managing data leakage in ML pipelines [23].

Furthermore, the use of the F1-score as the objective function in the tuning process reinforces the model's focus on the minority class. The combination of this anti-leakage pipeline and F1-score-based optimization demonstrates that the achieved performance improvements are not an artifact of methodological errors, but rather the result of a valid and controlled experimental design. Therefore, this section strongly supports Contribution, which is improving validity and trustworthiness in evaluating machine learning models on imbalanced data.

3.5. Discussion

3.5.1. Why the Proposed Stacking Model Performs Effectively

The proposed stacking architecture achieved strong performance because it combines complementary learning behaviors from XGBoost, LightGBM, and CatBoost within a unified meta-learning framework. Each boosting algorithm possesses different optimization characteristics and tree construction strategies, enabling the ensemble to capture heterogeneous nonlinear relationships in water quality data. To clarify the explainability structure of the proposed framework, SHAP analysis was conducted

separately for each boosting base learner (XGBoost, LightGBM, and CatBoost) and for the stacking meta-model based on Random Forest with passthrough=True. This separation is important because the SHAP explanations generated from the base learners represent how each boosting algorithm independently understands physicochemical water-quality characteristics, whereas the SHAP explanation generated from the stacking meta-model represents how predictive signals from multiple learners are fused into the final decision. The SHAP explanation of the XGBoost model, shown in Figure 4(a), indicates that the model relies heavily on strong nonlinear interactions among several dominant physicochemical features. Aluminium appears as the most influential contributor with a large negative SHAP value (−1.82), followed by chloramine, cadmium, lead, uranium, bacteria, and copper. The magnitude of these SHAP contributions suggests that XGBoost tends to construct sharper decision boundaries and aggressively separates unsafe and safe classes based on extreme feature patterns. This behavior explains why XGBoost contributes strong discriminative capability within the stacking architecture, particularly for samples with clear abnormal chemical characteristics.

A slightly different interpretation pattern is observed in the LightGBM model in Figure 4(b). Although aluminium and cadmium remain dominant features, the SHAP contributions are evenly distributed across multiple variables, including uranium, chloramine, viruses, lead, bacteria, and copper. Compared with XGBoost, LightGBM demonstrates a more adaptive interpretation pattern because it captures broader relationships among features through its leaf-wise tree growth mechanism. This indicates that LightGBM contributes complementary information by learning more distributed feature interactions rather than relying only on highly dominant variables.

Meanwhile, the CatBoost explanation presented in Figure 4(c) reveals another interpretation characteristic. CatBoost still identifies aluminium and cadmium as important contributors, but the SHAP contribution magnitudes appear smoother and more stable compared with XGBoost. Additional features such as silver, perchlorate, nitrates, and radium also contribute to the prediction process. This indicates that CatBoost is capable of capturing subtle dependencies among physicochemical variables and provides stable interpretation patterns even when feature interactions become more complex. Such behavior is beneficial in ensemble learning because it reduces prediction instability caused by highly fluctuating feature contributions.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

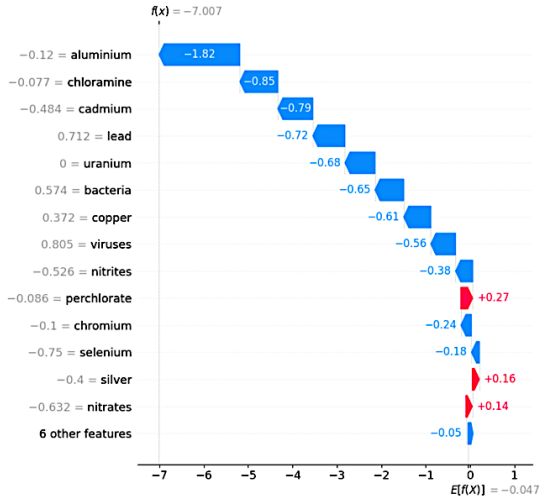

Figure (4a). The SHAP of the XGBoost

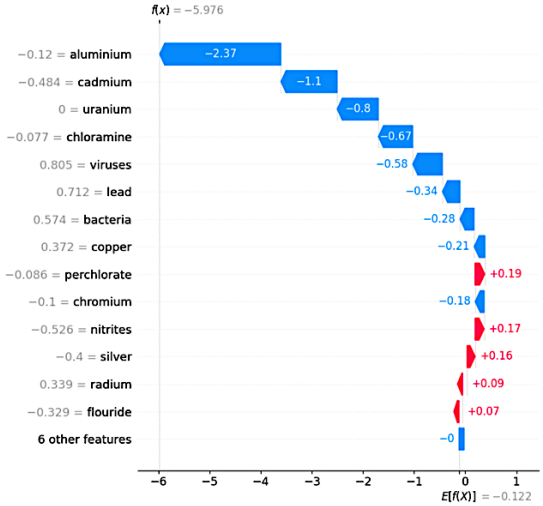

Figure (4b). The SHAP of the LGBoost

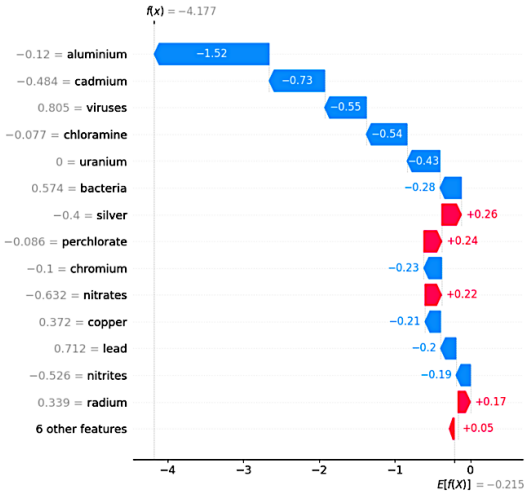

Figure (4c). The SHAP of the CatBoost meta-

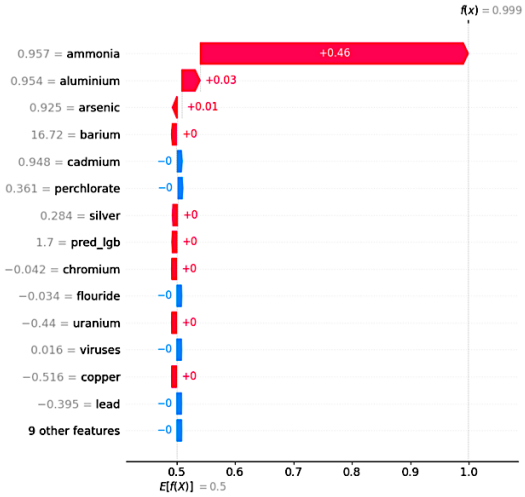

Figure (4d). The SHAP of the Stacking model

Although the three boosting models exhibit different SHAP contribution structures, they consistently identify chemically relevant parameters such as aluminium, cadmium, chloramine, uranium, and bacteria as important determinants of water quality. This consistency strengthens the reliability of the ensemble because the models learn overlapping yet complementary representations of unsafe water conditions. Therefore, the stacking framework benefits not only from predictive diversity but also from complementary feature understanding across multiple boosting algorithms.

The SHAP explanation of the stacking meta-model, shown in Figure 4(d), provides a substantially different interpretation perspective. Unlike the base learners, the stacking meta-model does not rely solely on raw physicochemical features. Because the proposed framework uses passthrough=True, the meta-model receives both the original features and the prediction outputs from the base learners as meta-features. Consequently, the SHAP explanation at the stacking level reflects how the framework performs adaptive signal fusion between original water-quality attributes and predictive confidence generated by the boosting models.

Figure 4(d) shows that ammonia becomes the most dominant feature in the stacking meta-model with a SHAP contribution of +0.46, which is substantially larger than the contributions of other variables. This finding differs from the base learners, where aluminium consistently appeared as the strongest contributor. The shift of dominance from aluminium in the base models to ammonia in the stacking model indicates that the meta-model learns which features are globally the most reliable after integrating predictions from multiple learners. In addition, the appearance of pred_lgb as one of the influential meta-features confirms that the stacking model explicitly utilizes predictive signals produced by LightGBM when constructing the final decision boundary.

This result demonstrates that the proposed stacking architecture is not merely performing majority voting or simple averaging. Instead, the meta-model performs explainable signal fusion by learning when to trust original physicochemical variables and when to rely on predictive tendencies from the base learners. The inclusion of pred_lgb as a relevant SHAP feature indicates that LightGBM contributes useful discriminative information to the final classification process, while the dominance of ammonia suggests that this feature becomes the most stable global indicator after ensemble integration.

3.5.2. Explainable AI (SHAP) for Local Interpretation: Audit of TP/FN/FP Cases and Assess Their Benefits

To complement the global interpretation, this study also conducted a SHAP-based local interpretation analysis on True Positive (TP), False Negative (FN), and False Positive (FP) cases. This approach allows model evaluation at the individual level, providing deeper insight into model behaviour.

FN (False Negative), where unsafe water is predicted as safe. This case is of primary interest because it represents a situation where unsafe water is predicted as safe, representing the most critical error. Local analysis, such as the SHAP waterfall visualization for the False Negative (FN) case in Figure 6, shows how a combination of features causes the model to incorrectly classify unsafe water as safe. This indicates that this error occurs due to a combination of offsetting features, where some critical parameters that indicate an unsafe condition are masked by other features within the normal range. This indicates that the model error is not random, but rather the result of complex interactions between features. Figure 5 shows that the ammonia feature makes a significant negative contribution (-0.27) to the prediction, pushing the model toward the safe class (f(x) ≈ 0.199), even though the actual condition is unsafe. This suggests that the ammonia values in these samples are within a range that is not extreme enough to trigger classification as unsafe, or that there are other features that offset their influence.

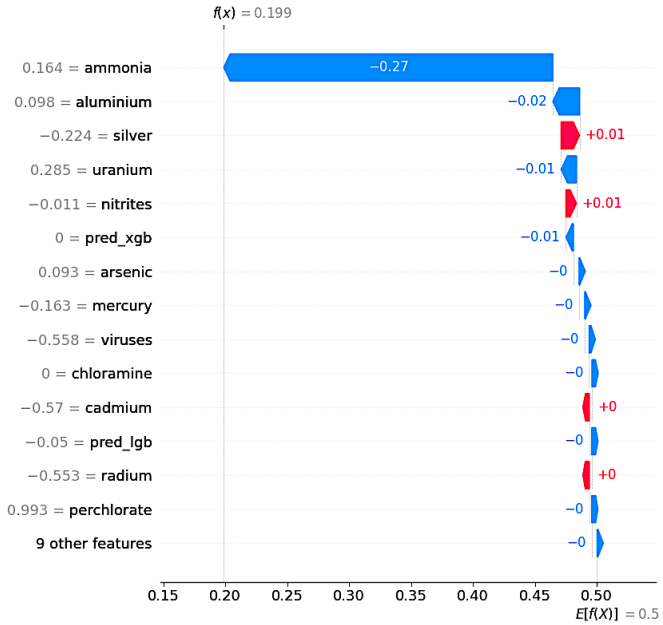

Figure 5. SHAP Waterfall Plot – False Negative (FN) Case

FP (False Positive) provides insight into operational costs: if the model marks safe as unsafe too often, there will be unnecessary retesting/mitigation consequences. Local SHAP helps understand the triggers so that thresholds or policies can be set more realistically. In the SHAP waterfall visualization for the False Positive (FP) case, showing the contribution of features that cause the model to classify safe water as unsafe. Figure

6 shows that a combination of several features such as ammonia (+0.07), arsenic (+0.02), and barium (+0.02) pushes the prediction towards unsafe (f(x) ≈ 0.522), even though the actual condition is safe. This indicates that the model tends to be sensitive to small changes in several parameters at once.

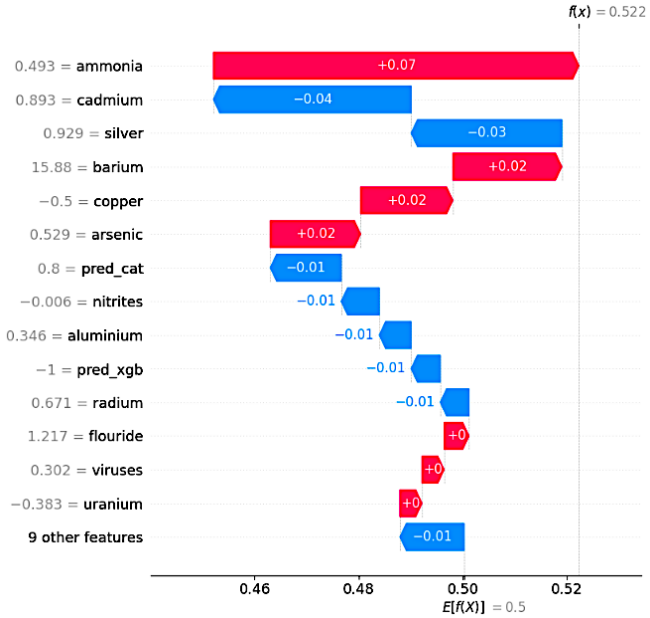

Figure 6. SHAP Waterfall Plot – False Positive (FP) Case

This analysis shows that model errors are not random, but rather the result of complex interactions between features, which can only be revealed through local interpretation. This approach offers practical benefits in the form of risk-oriented debugging, where users can understand the specific conditions that lead to prediction errors. TP (True Positive) helps ensure that the reason for the "unsafe" prediction is consistent with the dominant global feature (e.g., high ammonia), so that global and local interpretations reinforce each other, rather than contradict each other. The SHAP waterfall visualization for the True Positive (TP) case in Figure 7 below shows the contribution of each feature in driving the prediction toward the "unsafe" class.

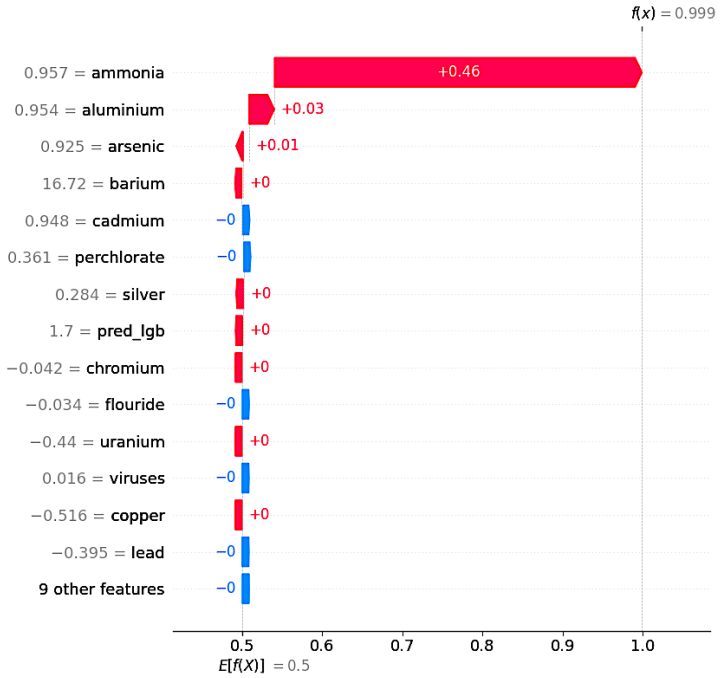

Figure 7. SHAP Waterfall Plot – True Positive (TP) Case

Figure 7 shows that the ammonia feature makes the largest positive contribution (+0.46) to the increased probability of the unsafe class, followed by aluminum and arsenic, which have smaller contributions. The final prediction value (f(x) ≈ 0.999) indicates a very high level of model confidence. This interpretation indicates the model's decisions are not black-box, but rather based on a combination of logically relevant features. Consistency between ammonia's dominance in the global analysis (Figure 6) and local contributions (Figure 7) indicates that the model has interpretive stability, which strengthens Contribution this research, which focuses on global-local explainability. The SHAP approach to assess dominant parameter contributors and relate model outputs to factor interpretations has also been used in recent XAI studies in the water quality/environmental domain, emphasizing the importance of transparency of key parameters and validation of interpretations [21], [24], [25].

4. CONCLUSION

This study proposed an ensemble stacking framework for imbalanced water quality classification by integrating XGBoost, LightGBM, and CatBoost with a Random Forest meta-learner using a passthrough mechanism within an anti-leakage pipeline consisting

of RobustScaler, SMOTE, and cross-validation, while Explainable AI based on SHAP was applied to provide both global and local model interpretations. The experimental results showed that the proposed framework achieved an F1-score of 0.8563 for the minority class and a ROC-AUC of 0.9846, indicating promising classification performance and balanced precision–recall under imbalanced conditions. SHAP analysis consistently identified ammonia as the dominant feature influencing water quality prediction, while local explanations on True Positive, False Negative, and False Positive cases demonstrated that prediction outcomes were driven by specific feature interactions rather than random model behavior. The findings also indicate that the stacking architecture benefits from complementary feature understanding across multiple boosting models and from adaptive signal fusion at the meta-model level. Therefore, the proposed framework shows potential as a more transparent and reliable approach for risk-aware water quality monitoring and other imbalanced classification problems. However, because this study was conducted using a single dataset and limited model configurations, further validation using broader datasets, additional baseline comparisons, alternative ensemble strategies, and real-world deployment scenarios is still necessary to strengthen the generalizability and robustness of the proposed approach.

REFERENCES

[1] Ms. M. Nandhini, "Water Quality Prediction Using Machine Learning Technique," *IJIREEICE*, vol. 13, no. 12, Dec. 2025, doi: 10.17148/IJIREEICE.2025.131206.

[2] W. Chen, D. Xu, B. Pan, Y. Zhao, and Y. Song, "Machine Learning-Based Water Quality Classification Assessment," *Water (Basel)*, vol. 16, no. 20, p. 2951, Oct. 2024, doi: 10.3390/w16202951.

[3] S. Yadav and G. P. Bhole, "Handling Imbalanced Dataset Classification in Machine Learning," in *2020 IEEE Pune Section International Conference (PuneCon)*, IEEE, Dec. 2020, pp. 38–43. doi: 10.1109/PuneCon50868.2020.9362471.

[4] L. Zhang and D. Jánošík, "Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches," *Expert Syst. Appl.*, vol. 241, p. 122686, May 2024, doi: 10.1016/j.eswa.2023.122686.

[5] R. Rivaldo, R. Taufik, I. S. Ilman, and O. D. E. Wulansari, "A Comparative Study of XGBoost, LightGBM, and CatBoost Models for Customer Churn Prediction in the Banking Industry," *Jurnal Pepadun*, vol. 6, no. 2, pp. 178–187, Aug. 2025, doi: 10.23960/pepadun.v6i2.277.

[6] R. Liu *et al.*, "Stacking Ensemble Method for Gestational Diabetes Mellitus Prediction in Chinese Pregnant Women: A Prospective Cohort Study," *J. Healthc. Eng.*, vol. 2022, pp. 1–14, Sep. 2022, doi: 10.1155/2022/8948082.

[7] M. Munsarif, M. Sam'an, and S. Safuan, "Peer to peer lending risk analysis based on embedded technique and stacking ensemble learning," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 6, pp. 3483–3489, Dec. 2022, doi: 10.11591/eei.v11i6.3927.

[8] P. Netayawijit, W. Chansanam, and K. Sorn-In, "Interpretable Machine Learning Framework for Diabetes Prediction: Integrating SMOTE Balancing with SHAP Explainability for Clinical Decision Support," *Healthcare*, vol. 13, no. 20, p. 2588, Oct. 2025, doi: 10.3390/healthcare13202588.

[9] P. Lakkarasu, "Designing and deploying scalable MLOps pipelines for continuous artificial intelligence model training and delivery," in *Designing Scalable and Intelligent Cloud Architectures: An End-to-End Guide to AI Driven Platforms, MLOps Pipelines, and Data Engineering for Digital Transformation*, Deep Science Publishing, 2025, pp. 28–42. doi: 10.70593/978-93-49910-08-9_3.

[10] S. Yang, Z. Huang, W. Xiao, and X. Shen, "Interpretable Credit Default Prediction with Ensemble Learning and SHAP," in *2025 International Conference on Artificial Intelligence, Human-Computer Interaction and Natural Language Processing (ICAHN)*, IEEE, May 2025, pp. 102–106. doi: 10.1109/ICAHN67688.2025.00027.

[11] O. Mermer, E. Zhang, and I. Demir, "A Comparative Study of Ensemble Machine Learning and Explainable AI for Predicting Harmful Algal Blooms," *Big Data and Cognitive Computing*, vol. 9, no. 5, p. 138, May 2025, doi: 10.3390/bdcc9050138.

[12] N. G. Rezk, S. Alshathri, A. Sayed, and E. El-Din Hemdan, "EWAIS: An Ensemble Learning and Explainable AI Approach for Water Quality Classification Toward IoT-Enabled Systems," *Processes*, vol. 12, no. 12, p. 2771, Dec. 2024, doi: 10.3390/pr12122771.

[13] Z. B. Tadese *et al.*, "Interpretable prediction of acute respiratory infection disease among under-five children in Ethiopia using ensemble machine learning and Shapley additive explanations (SHAP)," *Digit. Health*, vol. 10, Jan. 2024, doi: 10.1177/20552076241272739.

[14] N. Nasir *et al.*, "Water quality classification using machine learning algorithms," *Journal of Water Process Engineering*, vol. 48, p. 102920, Aug. 2022, doi: 10.1016/j.jwpe.2022.102920.

[15] "Exploring The Effectiveness Of Different Data Cleaning Techniques For Improving Data Quality in Machine Learning," *Humanitarian and Natural Sciences Journal*, vol. 4, no. 7, Jul. 2023, doi: 10.53796/hnsj4711.

[16] Prof. Arati K Kale and Dr. Dev Ras Pandey, "Data Pre-Processing Technique for Enhancing Healthcare Data Quality Using Artificial Intelligence," *Int. J. Sci. Res. Sci. Technol.*, pp. 299–309, Jan. 2024, doi: 10.32628/IJSRST52411130.

[17] A. A. Soomro *et al.*, "Data augmentation using SMOTE technique: Application for prediction of burst pressure of hydrocarbons pipeline using supervised machine learning models," *Results in Engineering*, vol. 24, p. 103233, Dec. 2024, doi: 10.1016/j.rineng.2024.103233.

[18] X. Ye, W. Xu, X. Ye, D. Long, Q. Yin, and B. Huang, "Stroke Prediction Using the Trust Evaluation with Data Leakage Avoiding," *J. Phys. Conf. Ser.*, vol. 2560, no. 1, p. 012051, Aug. 2023, doi: 10.1088/1742-6596/2560/1/012051.

[19] N. Rathnayake, T. Linh Dang, and Y. Hoshino, "Designing and Implementation of Novel Ensemble model based on ANFIS and Gradient Boosting methods for Hand Gestures Classification," in *The 11th International Symposium on Information and Communication Technology*, New York, NY, USA: ACM, Dec. 2022, pp. 283–289. doi: 10.1145/3568562.3568598.

[20] Z. Chen, "The Principle of Tree Explainer and Its Associated Validation," in *Proceedings of the 5th International Conference on Computer Information and Big Data Applications*, New York, NY, USA: ACM, Apr. 2024, pp. 1155–1162. doi: 10.1145/3671151.3671352.

[21] G. Zhao *et al.*, "Enhancing interpretability of tree-based models for downstream salinity prediction: Decomposing feature importance using the Shapley additive explanation approach," *Results in Engineering*, vol. 23, p. 102373, Sep. 2024, doi: 10.1016/j.rineng.2024.102373.

[22] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to <scp>SHAP</scp> analysis: Explaining supervised machine learning model predictions in drug development," *Clin. Transl. Sci.*, vol. 17, no. 11, Nov. 2024, doi: 10.1111/cts.70056.

[23] A. M. AbdulAbbas, R. Alkanany, Y. A. K. Al-Nuaimi, and Z. M. A. Al-Hamdawee, "A Sequential Data Preprocessing Pipeline for Diabetes Prediction: A Data Leakage Prevention and Dual-Validation Approach," *Engineering, Technology & Applied Science Research*, vol. 15, no. 6, pp. 30059–30066, Dec. 2025, doi: 10.48084/etasr.14155.

[24] R. K. Makumbura *et al.*, "Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like shapley additive explanations (SHAP) for interpreting the black-box nature," *Results in Engineering*, vol. 23, p. 102831, Sep. 2024, doi: 10.1016/j.rineng.2024.102831.

[25] A. Aldrees, M. Khan, A. T. B. Taha, and M. Ali, "Evaluation of water quality indexes with novel machine learning and SHapley Additive ExPlanation (SHAP) approaches," *Journal of Water Process Engineering*, vol. 58, p. 104789, Feb. 2024, doi: 10.1016/j.jwpe.2024.104789.