Erciyes University Journal of Institue Of Science and Technology

Research Article
10.65520/erciyesfen.1994336

Imprint:
Volume: 42(3)
Year: 2026
Page: 137-160

Halil İbrahim SARI[a]
Kübra KESER[b*]

[a] Master's Student, Kutahya Dumlupınar University, 43halilsari@gmail.com
[b] Asst. Prof., Kutahya Dumlupınar University, kubra.keser@dpu.edu.tr.

[*] Corresponding Author

Received: 7/14/2026
Accepted: 9/11/2026

Citation:
Halil İbrahim SARI, Kübra KESER (2026). Classification of Synthetic Biosensor Data with Quantum Simulations Using Feature Optimization and Explainable Machine Learning Supported by RFECV. *Erciyes University Journal of Institue Of Science and Technology*, 42(3), 137-160. https://doi.org/10.65520/erciyesfen.*1994336*

Screened by iThenticate for Authors & Researchers

Classification of Synthetic Biosensor Data with Quantum Simulations Using Feature Optimization and Explainable Machine Learning Supported by RFECV

Abstract

Background/Aim: Accurate detection of trace amounts of heavy metal toxins in environmental samples requires sensitive and reliable diagnostic systems. This study presents a machine learning framework to classify synthetic biosensor signals—generated via quantum simulation—representing five toxin classes: Control, Lead, Mercury, Cadmium, and Arsenic.

Methods: Methods: A total of 75 pre-calculated temporal, spectral, entropy, nonlinear, and wavelet features were utilized. To mitigate the risk of data leakage and obtain reliable performance estimates, the Recursive Feature Elimination with Cross-Validation (RFECV) method was integrated into a GroupKFold-based validation framework. The number of features selected across the five outer folds was 47, 73, 55, 47, and 41, respectively, resulting in an average of 52.6 features. Six ensemble learning algorithms were evaluated using a fixed 47-feature structure. Model interpretability was examined via SHAP analysis, while the contributions of feature groups were assessed through ablation and Wilcoxon signed-rank analyses.

Results: Results: Configurations with 75 and 41 features were evaluated in a separate direct comparison experiment, yielding average accuracies of 84.06% and 83.73%, respectively. The main evaluation was conducted on the 47-feature configuration. Among the six ensemble learning algorithms, HistGradientBoosting demonstrated the highest performance, achieving an average accuracy of 85.27% and a weighted F1-score of 85.25%. The Random Forest model exhibited the highest discriminative performance, with a Macro-AUC of 0.9799. The Friedman test indicated that performance differences among the algorithms were statistically significant ($p = 0.0186$), while Nemenyi post-hoc analysis assessed the relative performance differences between the models. The 47-feature configuration was used for the ablation analysis. The most significant drop in performance occurred upon the removal of frequency-domain features, followed by the removal of non-linear features. The Wilcoxon signed-rank test, applied to evaluate the contribution of non-linear features, demonstrated that this group provided a positive trend and contribution to model performance ($p = 0.0625$).

Conclusion: SHAP-based analyses have demonstrated that features such as Hjorth Mobility, D1 Energy, D1 Standard Deviation, Petrosian Fractal Dimension, and Spectral Entropy contribute significantly to the model's decisions. The results indicate that the combined use of GroupKFold-based validation, RFECV, and explainable AI methods facilitates the reliable and interpretable classification of biosensor data.

Keywords: Toxin, classification, biosensor, RFECV, XAI.

Except where otherwise noted, content in this article is licensed under a Creative Commons 4.0 International license. Icons by Font Awesome.

Öznitelik Optimizasyonu ve RFECV Destekli Açıklanabilir Makine Öğrenmesi ile Kuantum Simülasyonlu Sentetik Biyosensör Verilerinin Sınıflandırılması

Öz

Çevresel örneklerdeki eser miktardaki ağır metal toksinlerinin doğru bir şekilde tespiti, yüksek hassasiyete sahip teşhis sistemleri gerektirir. Kuantum hesaplama yeni öznitelik temsil yetenekleri sunsa da, fiziksel kuantum biyosensörleri henüz büyük ölçüde deneysel aşamadadır. Bu çalışmada, üretilen kuantum simülasyonlu sentetik biyosensör sinyallerini sınıflandırmak amacıyla tasarlanmış bir makine öğrenmesi mimarisi sunulmaktadır. Beş hedef toksin sınıfı (Kontrol, Kurşun, Cıva, Kadmiyum, Arsenik) genelindeki sinyal dinamiklerini temsil etmek üzere toplam 75 önceden hesaplanmış zamansal, spektral, entropi, doğrusal olmayan ve dalgacık özniteliği kullanılmıştır.

Veri sızıntısı riskini azaltmak ve daha güvenilir performans tahminleri elde etmek amacıyla, RFECV tabanlı öznitelik seçimi GroupKFold tabanlı çapraz doğrulama çerçevesinde gerçekleştirilmiştir. Beş dış katmanda RFECV tarafından seçilen öznitelik sayıları 47, 73, 55, 47 ve 41 olup ortalama 52.6 öznitelik olarak bulunmuştur. Ayrıca, 75 ve 41 öznitelikli yapılar ayrı bir doğrudan karşılaştırma deneyinde değerlendirilmiş ve ortalama doğruluk sırasıyla %84.06 ve %83.73 olarak elde edilmiştir. Ana değerlendirme 47 öznitelikli yapı üzerinde gerçekleştirilmiştir. Altı ensemble öğrenme algoritması arasında HistGradientBoosting %85.27 ortalama doğruluk ve %85.25 Weighted F1 skoru ile en yüksek performansı sergilemiştir. Rastgele Orman modeli 0.9799 Macro-AUC ile en yüksek ayırt etme performansını göstermiştir. Friedman testi algoritmalar arasındaki performans farklarının istatistiksel olarak anlamlı olduğunu göstermiş ($p = 0.0186$), Nemenyi post-hoc analizi ise modeller arasındaki göreli performans farklılıklarını değerlendirmiştir. Ablasyon analizinde 47 öznitelikli yapı kullanılmıştır. En belirgin performans düşüşü frekans alanı özniteliklerinin çıkarılmasıyla meydana gelmiş, bunu doğrusal olmayan özniteliklerin çıkarılması takip etmiştir. Doğrusal olmayan özniteliklerin katkısını değerlendirmek için uygulanan Wilcoxon signed-rank testi, bu grubun model performansına olumlu bir eğilim ve katkı sağladığını göstermiştir ($p=0.0625$). SHAP tabanlı analizler, Hjorth Mobility, D1 Energy, D1 Standard Deviation, Petrosian Fractal Dimension ve Spectral Entropy gibi özniteliklerin model kararlarına önemli katkılar sağladığını göstermiştir. Sonuçlar, GroupKFold tabanlı doğrulama, RFECV ve açıklanabilir yapay zekâ yöntemlerinin birlikte kullanımının biyosensör verilerinin güvenilir ve yorumlanabilir sınıflandırılmasına katkı sağladığını göstermektedir.

Anahtar kelimeler: Toksin, sınıflandırma, biyosensör, RFECV, XAI.

1. INTRODUCTION

Environmental toxins are among the significant pollutants that can have long-term effects on human health and ecosystems. Heavy metals, particularly lead, mercury, cadmium, and arsenic, can cause irreversible damage to many biological structures, especially the nervous system, kidneys, and cardiovascular system [1]. Therefore, early and reliable accurate identification of environmental toxins is considered a critical requirement for protecting public health and managing environmental risks [2].

Currently, laboratory-based analytical methods such as Inductively Coupled Plasma Mass Spectrometry (ICP-MS), Atomic Absorption Spectroscopy (AAS), and High-Performance Liquid Chromatography (HPLC) are considered the gold standard for the determination of environmental toxins. While these traditional approaches offer high selectivity and accuracy even at ultra-low concentrations, they have significant limitations in real-time field applications. Traditional laboratory-based analytical methods have several limitations in real-time applications due to their high cost, the need for expert personnel, and the time-consuming nature of the analytical process [2,3].

New approaches to the rapid and accurate identification of environmental toxins have been developed as a consequence of recent advances in biosensor technologies. AI-based biosensor systems are especially attracting attention due to their high accuracy and rapid decision-making capabilities in the analysis of complex biological data. In this study, we tried to solve the problem of biosensors' inability to distinguish multiple toxins in complex environmental settings using classical machine learning, by integrating Microbial Electrochemical Systems (MECs) with different algorithms. This

CC: BY-NC-ND

successfully separated complex current-time signals generated by different toxins simultaneously present in water sources, achieving high accuracy [4]. A Convolution Neural Network (CNN) model was used to process information from sensors for the detection of toxic substances such as cadmium (Cd2+) and chromium (Cr6+) in water [5]. A machine learning-based biosensor calibration model has enabled the real-time and much more accurate field detection of the cyanobacterial toxin microcystin (MC-LR) in various aquatic environments [6]. However, the performance of classical machine learning approaches can be influenced by the complexity, dimensionality, and redundancy of biosensor signal representations, particularly in multi-class toxin classification tasks.

Quantum Machine Learning (QML) and quantum-enhanced neural networks exploit quantum computing principles such as superposition and entanglement to explore alternative computational and feature-representation strategies for complex learning problems [7]. Advances in quantum computing have opened new possibilities for biosensor research, including the generation and analysis of Quantum-Simulated Synthetic Biosensor Data. These systems offer researchers a flexible and cost-effective working environment, particularly in the processes of synthetic data generation and model development. Thus, in situations where real experimental data is limited, synthetic datasets constitute a significant alternative for the development, comparison, and validation of algorithms. Collecting real-time experimental data is costly, time-consuming, and sometimes dangerous. Recently, quantum-based generative approaches have also been investigated for generating synthetic datasets from limited experimental data, providing controlled environments for model development and algorithm evaluation. This makes model training cost-effective while facilitating algorithm validation. In the literature, diamond nanoparticles with Nitrogen-Vacuum (NV) centers are used as ultra-sensitive quantum biosensors at the cellular level and in microenvironments. It is stated that the surface instability problem of quantum biosensors is overcome through design and coherence times are increased by up to 3.5 times [8].

Recent advances in biosensor technologies have significantly improved the sensitivity and selectivity of environmental toxin detection. However, the size and complexity of biosensor signals can limit the predictive performance of machine learning and deep learning models. Recent studies are increasingly focusing on the integration of feature selection and Explainable Artificial Intelligence (XAI) techniques to improve both classification accuracy and model interpretability [9,10]. Feature selection has become a crucial preprocessing step in machine learning-based biosensor applications because it eliminates redundant and irrelevant variables, reduces computational complexity, and minimizes overfitting while maintaining prediction performance. Among the wrapper-based feature selection approaches, RFECV has garnered significant attention because it automatically identifies the optimal subset of features by recursively eliminating less informative variables based on model performance. RFECV uses cross-validation to identify the optimal number of conserved features, thereby improving prediction models [11,12]. Several recent studies have demonstrated the effectiveness of RFECV in various machine learning applications. For example, an explainable respiratory toxicity prediction framework has shown that combining correlation-based filtering, RFECV, and multiple machine learning classifiers significantly reduces feature dimensionality while increasing prediction accuracy. Furthermore, SHAP analysis has been integrated to interpret the contribution of selected molecular descriptors, highlighting the complementary roles of feature optimization and explainability [11]. Similarly, an explainable machine learning framework for water quality management has been proposed by integrating RFECV with explainable machine learning. It is stated that RFECV reduces the input variable size by approximately 69-85% without sacrificing comparable accuracy, highlighting the effectiveness of cross-validated feature selection for environmental monitoring applications[12].

Beyond prediction performance, model interpretability has become a necessity. Traditional black-box machine learning algorithms often provide high classification accuracy but fail to explain how prediction decisions are generated, reducing user confidence in practical applications. XAI overcomes this limitation by revealing the relationships between input variables and model outputs, thereby increasing transparency and facilitating scientific interpretation. Recent reviews have highlighted the increasing importance of XAI in many fields due to the need for interpretable decision support

CC: BY-NC-ND

systems[10,13]. Furthermore, recent environmental studies have shown that SHAP not only improves model transparency but also provides insights into pollutant behavior, adsorption mechanisms, and environmental remediation processes[14,15].

Despite these advances, there remains a need for integrated frameworks that combine cross-validated feature optimization, source- cross-validation folds validation, ensemble learning, and model-specific explainability for the classification of quantum-simulated biosensor signals. In particular, the contribution of different signal-feature domains and the effect of RFECV-based dimensionality reduction should be evaluated systematically rather than relying solely on classifier-level performance comparisons.

To address the challenges associated with limited experimental data and high-dimensional biosensor signal representations, this study aimed to develop a reliable XAI-supported classification approach using quantum-simulated synthetic biosensor data. For this purpose, the 75 signal features provided by the dataset were used as the initial feature space, and RFECV was employed to identify a compact feature subset. The original 75-feature representation and the RFECV-selected subset were directly compared using identical five-fold GroupKFold splits and evaluation metrics to assess the effect of feature optimization on predictive performance and computational cost. To prevent data leakage that may arise from the coexistence of similar signal windows originating from the same raw sensor source in the training and test sets, the Source_File information was used as a group label, and a GroupKFold-based source cross-validation folds validation strategy was applied throughout the experiments. Histogram-based Gradient Boosting (HistGradientBoosting), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Extra Trees, and Categorical Boosting (CatBoost) algorithms were comparatively evaluated. Model performance was assessed using accuracy, macro-averaged precision, recall and F1-score, as well as Matthews Correlation Coefficient (MCC), Cohen's Kappa, and Receiver Operating Characteristic (ROC)-Area Under the Curve (AUC). Furthermore, the statistical significance of performance differences between the models was examined using the Friedman and Nemenyi post-hoc tests. SHAP-based explainable artificial intelligence analysis was performed to improve the interpretability of model decisions, while ablation experiments were conducted to evaluate the contribution of different feature domains to classification performance.

2. MATERIALS AND METHODS

The open-access "Quantum-AI Enhanced Biosensor Dataset for Environmental Toxin Detection: A Synthetic Data Approach" dataset was used for analyses aimed at classifying environmental toxins[16–18]" dataset was used for analyses aimed at classifying environmental toxins. It consists of synthetic biosensor data generated using the Qiskit Aer simulator. The dataset contains 160.000 synthetic biosensor samples evenly distributed across 80 independent Source_File groups, each containing 2000 signal samples. The balanced class distribution enables an unbiased evaluation of classification model performance across classes (Table 1).

Signals were generated for five classes of environmental toxins (Mercury, Lead, Cadmium, Arsenic, and Control) under controlled simulation conditions. To prevent information leakage from related samples originating from the same simulation group, the Source_File variable was used as the grouping variable throughout the GroupKFold validation procedure. The dataset contains 75 features for each sample covering time-domain, frequency-domain, entropy, nonlinear, and wavelet-based characteristics. These features represent different properties of biosensor signals that can provide discriminative information between toxin classes. The main reasons we chose this dataset are its open accessibility, balanced class structure, and the fact that it provides a suitable framework for comparing different machine learning approaches on synthetic data derived from biosensor data. Furthermore, the dataset allows for the objective and reproducible evaluation of proposed methods in the field of environmental toxin classification.

In the later stages of the study, RFECV-based feature selection was performed on these original features, and the subset contributing most to the classification performance was identified. No additional hand-crafted feature extraction was performed in this study; instead, the original dataset-

CC: BY-NC-ND

provided feature set was optimized using RFECV. The RFECV procedure selected 41 features from the original space of 75 features. To directly assess the impact of feature selection, this subset of selected features was evaluated using the same five-fold GroupKFold partitioning and the same evaluation criteria.

The synthetic data method alleviates the cost, time and ethical limitations to acquire real experimental data, while offering a controlled and reproducible evaluation environment to develop and compare the machine learning algorithms.

Table 1. Key features and values of the dataset used.

Features	Value
Data type	Synthetic biosensor data
Data generation	Qiskit Aer Simulator
Total number of samples	160.000
Number of classes	5
Classes	Control, Lead, Mercury, Cadmium, Arsenic
Total attributes	75
Feature Domains	Time, Frequency, Nonlinear, Wavelet, Entropy
Source Groups	Source_File

The study was performed based on a multi-stage methodology including data preparation, pre-computed feature selection, feature selection, model training and performance evaluation, and finally XAI analysis (Figure 1). The following subsections detail the method, describing each step in detail.

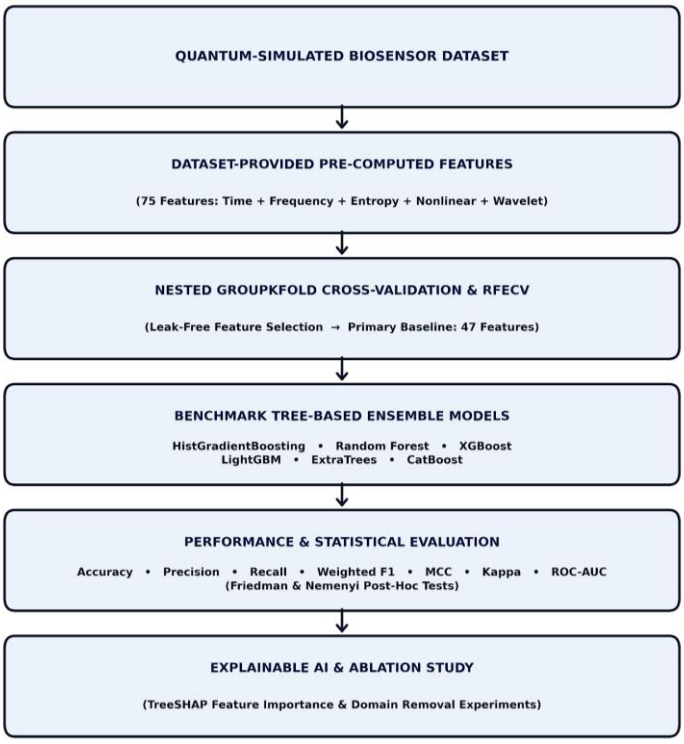


Figure 1. Proposed framework

2.1. Data Preprocessing

In an effort to achieve consistent and reliable results from the machine learning algorithms, the dataset was preprocessed through several steps before classification. The dataset was checked for

integrity, for missing and duplicate records and no missing or duplicate observations were found that could impact the analysis. Next, we analyzed the class distribution and verified the dataset for a balanced sample structure among five environmental toxin classes (Table 2).

Table 2. Operations performed during the data preprocessing process

Process	Explanation
Missing data check	The dataset contained no missing observations.
Repeated record check	No duplicate entries were detected.
Examining class distribution	The dataset was confirmed to be balanced across the five classes.
Label Encoding	The target variable has been converted into numerical labels.
Determining group information	The Source_File variable is used as group information for GroupKFold validation.
Data balancing	Not applied (dataset is balanced).
Normalization/Standardization	Not applied

The Source_File variable in the dataset represents the original raw sensor signal from which each sample is derived. Samples with the same Source_File tag are obtained from different time windows generated from the same raw signal. Therefore, there is a high degree of similarity between samples belonging to the same group. If these samples were randomly distributed into training and test sets, data leakage would occur, and the model performance would be overestimated. To prevent this problem, the GroupKFold cross-validation method was used during the evaluation phase. This prevented samples belonging to the same Source_File group from being present simultaneously in the training and test sets, and the model's true generalization performance was evaluated more reliably.

In this study, the environmental toxin classes used as target variables were encoded using Label Encoding to convert them into the numerical format required by scikit-learn-based classification algorithms. This process only provided numerical representation of the class labels and did not define any ranking relationships between the classes. To realistically and reliably evaluate model performance, the Source_File information in the dataset was used as a group variable. This prevented samples from the same source from appearing together in both the training and test sets, thus preventing potential data leakage. This approach formed the basis of the GroupKFold cross-validation strategy used in the study.

No sample adjustment method was applied due to the balanced class distribution of the dataset. Furthermore, because the tree-based machine learning algorithms used in the study are not scaling-sensitive, no normalization or standardization was performed on the features. After completing the preprocessing, the feature matrix (x) and target variable (y) were created, and the feature selection and classification stages were initiated.

2.2. Original Feature Set and Feature Domains

To fully characterize the different properties of biosensor signals for environmental toxins, the dataset provides 75 pre-computed features across multiple analytical domains. These features encompass time domain, frequency domain, entropy, nonlinear analysis, and Wavelet transform characteristics. This approach allows a richer feature space by taking into account statistical properties, spectral structure, dynamic behavior and time-frequency characteristics of the signals jointly.

Time domain features provide the basic statistical and amplitude features of the signal and frequency domain features provide the spectral distribution and dominant frequency features of the signal. The nonlinear features reveal the complex and dynamic structure of the signal and wavelet based features analyze the temporal variations of signal in different frequency bands at high resolution. So the different effects of the environmental toxins on the biosensor signals are studied

CC: BY-NC-ND

from all sides. Table 3 shows the feature groups used in the study and representative features of each group.

Table 3. Feature groups and representative features pre-computed from biosensor signals

Feature Group	Representative Features	Description
Time Domain	Mean, Median, Std, RMS, Peak, Peak_Amplitude, Max, Interquartile Range (IQR)	Statistical descriptors
Frequency Domain	Spectral Entropy, Spectral Centroid, Spectral Bandwidth, RollOff	Spectral characteristics
Entropy Domain	Shannon Entropy, Log-Energy Entropy, Rényi Entropy, Tsallis Entropy, Normalized Entropy	Signal uncertainty and complexity
Nonlinear	Hjorth Activity, Hjorth Mobility, Hjorth Complexity, Petrosian FD	Signal complexity
Wavelet	A1 Mean, A1 Std, A2 Mean, D1 Std, D2 Energy, D3 Std ...	Multi-resolution decomposition

2.3. Feature Selection Using RFECV

It is known that not all dataset-provided features contribute equally to classification performance. Including unnecessary or low-discrimination features in the model can increase computational cost and negatively affect the model's generalization ability. Therefore, the RFECV method was used to determine the features that contribute most to classification performance. RFECV evaluates model performance through cross-validation, gradually removing features with the lowest importance from the dataset in each iteration, and automatically determining the optimum feature set that provides the highest accuracy value [19–21]. In this study, RFECV was performed using the Random Forest classifier with accuracy as the evaluation criterion and a GroupKFold-based cross-validation strategy. Across the five outer folds, the number of features selected by RFECV was 47, 73, 55, 47, and 41, yielding an average of 52.6 features. The primary benchmark evaluation was established using the 47-feature configuration.

2.4. Classification Models

Six different ensemble learning algorithms were used to comparatively evaluate the classification success of features determined by the RFECV method. These are Random Forest, XGBoost, LightGBM, CatBoost, ET, and HistGradientBoosting models. The combined evaluation of these algorithms, which have different learning mechanisms, aims to reveal the generalizability of the proposed biosensor approach against different classifiers.

To reliably evaluate the performance of machine learning models, a 5-fold GroupKFold cross-validation method was used. In this method, when the dataset was divided into training and test sets, all samples within the same Source_File (same biosensor signal record) were kept in the same fold. The data leakage risk was removed by not allowing samples belonging to the same biosensor signal record to be present in both the training and test set. This approach allows the model to be evaluated on biosensor signal records it has not previously encountered, more accurately reflecting real-world application scenarios and enabling a more reliable measurement of the model's generalization ability. Furthermore, all classification algorithms were trained and tested on the same GroupKFold folds, ensuring a fair and comparable performance evaluation among the algorithms (Figure 2).

To ensure a fair and unbiased comparison among the evaluated classifiers, all models were trained and tested using the same experimental protocol and the same GroupKFold partitions. In this study, hyperparameter optimization was deliberately omitted because the primary objective was not to achieve the highest performance of individual algorithms, but rather to objectively evaluate the contribution of the proposed RFECV-based feature selection and feature engineering approach to different ensemble learning algorithms under the same learning conditions. This aimed to attribute the observed performance differences to the proposed feature extraction and selection strategy, rather

CC: BY-NC-ND

than to model-based hyperparameter adjustments.

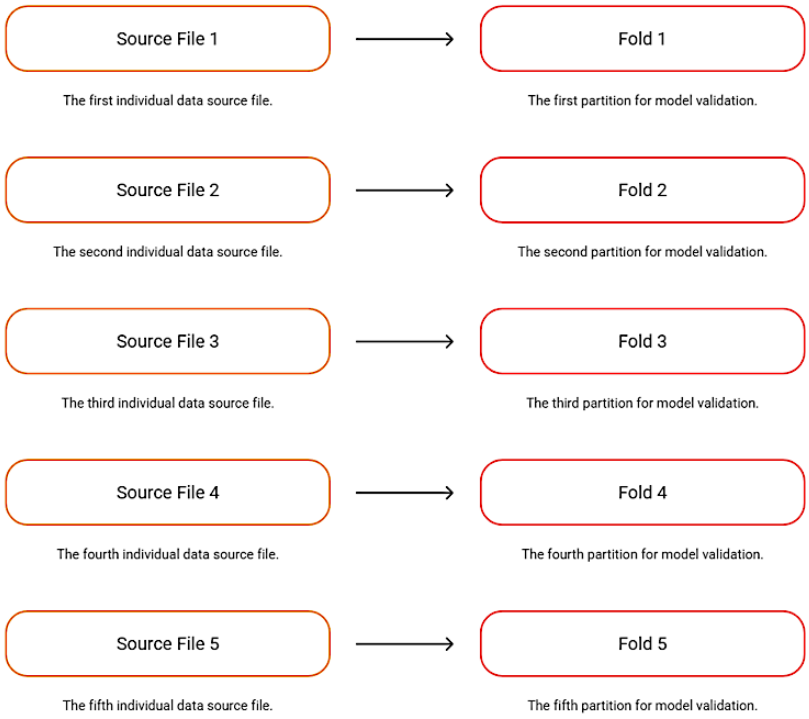


Figure 2. GroupKFold-based data partitioning strategy

All models were evaluated on the same training and testing bins, thus ensuring a fair performance comparison between the algorithms. The performance of the classification models was analyzed using seven different performance metrics to allow for a comprehensive evaluation from various perspectives (Table 4).

Accuracy represents the ratio of correctly classified samples to the total number of samples. Precision indicates how many of the positively predicted samples were correctly predicted, while Recall shows the rate at which true positive samples were correctly detected. The F1-score, as the harmonic mean of the Precision and Recall values, provides a more reliable assessment in unbalanced datasets. Since Accuracy alone may not fully reflect model performance, MCC and Cohen's Kappa coefficients, which take into account the balance between classes, were also calculated. In addition, the ROC-AUC value was used to evaluate the discrimination performance of the models at different decision thresholds. All performance metrics were calculated by averaging the test results obtained in the GroupKFold cross-validation process and were used in comparing different classification algorithms. Multi-class ROC analysis was calculated using the macro-average approach.

Table 4. Performance evaluation metrics used in this study.

Metric	Description
Accuracy	Overall classification accuracy
Precision	Correctness of positive predictions
Recall	Ability to identify positive samples
Weighted F1-score	Weighted harmonic mean of Precision and Recall
MCC	Balanced correlation between predictions and true labels
Cohen's Kappa	Agreement beyond chance
ROC-AUC	Overall discrimination capability of the classifier

Variability was expressed as Mean ± Standard Deviation across 5 external validation folds. For explainability, TreeSHAP was run across the test folds of the GroupKFold structure.

In addition to the main multi-model evaluation, a direct feature-set comparison was performed to quantify the effect of RFECV-based feature selection. Using the same classifier, identical five-fold GroupKFold partitions, and the same evaluation metrics, the original 75-feature configuration was compared with the RFECV-selected 41-feature configuration. This comparison was used to determine whether dimensionality reduction affected predictive performance and computational cost.

2.5. XAI

For machine learning models to not only achieve high classification success but also to be able to explain their decisions, this is crucial for the development of reliable artificial intelligence systems. Therefore, XAI methods were used to interpret the decision-making mechanism of the proposed model. In the first stage, feature importance scores obtained from Random Forest, XGBoost, and LightGBM models were calculated and normalized, and a common Consensus Feature importance analysis was performed. For the HistGradientBoosting model, the contribution of features to classification success was evaluated using the permutation importance method. Finally, SHAP analysis was applied to interpret both the overall behavior of the model and individual predictions. To conduct a more detailed evaluation, the SHAP method was used. SHAP analysis quantitatively reveals the contribution of each feature to the model predictions and allows for the interpretation of both the overall behavior of the model and individual predictions. In this way, the proposed approach not only provides a high-performance classification model but also increases the interpretability of the results by explaining to what extent the model decisions are affected by which features.

All experiments were conducted using Python 3.12.13 on the Google Colab environment. The implementation relied on NumPy 2.0.2, Pandas 2.2.2, Scikit-learn 1.6.1, XGBoost 3.3.0, LightGBM 4.6.0, and CatBoost 1.2.10. To ensure reproducibility, a fixed random seed (random_state = 42) was consistently applied to all data partitioning procedures, RFECV feature selection, and machine learning models. Model evaluation was performed using GroupKFold cross-validation under identical experimental conditions, and both training and inference times were recorded for all benchmark models.

2.6 Ablation Study

Ablation experiments were conducted to examine the contribution of different feature groups to classification performance. The primary 47-feature configuration was used as the reference baseline model, and the contributions of the time-domain, frequency-domain, entropy-domain, nonlinear, and wavelet-based feature groups were evaluated using the HistGradientBoosting model. For each ablation condition, the corresponding feature group was excluded while the remaining features were retained, and the model was retrained under the same evaluation protocol. The resulting performance changes were compared with the complete feature-set configuration to quantify the contribution of the evaluated feature groups to classification performance.

2.7. Statistical Analysis

Considering that comparing the performance of different machine learning algorithms solely numerically might be insufficient, the statistical significance of the observed performance differences was also evaluated. A non-parametric Friedman test was applied to compare the performance of six classification algorithms across five GroupKFold cross-validation folds. When the Friedman test showed a statistically significant difference, a Nemenyi post-hoc test was applied to examine pairwise differences between the algorithms. In addition, the mean ranking of each algorithm was calculated to summarize its relative performance across the cross-validation folds. The significance level was set at $p < 0.05$ for all statistical analyses.

3.RESULTS AND DISCUSSION

This section systematically presents and evaluates the experimental findings related to the

CC: BY-NC-ND

proposed environmental toxin classification framework. First, the structure of the generated feature set and the findings obtained from RFECV-based feature selection are discussed. Then, the performance of the proposed approach is compared with reference ensemble learning algorithms using various performance metrics. Following this, the model's decision-making mechanism is interpreted using SHAP-based XAI analysis, and the contribution of different feature groups to classification success is examined within the scope of ablation studies. SHAP analysis was used to provide a model-specific description of feature contributions in the trained XGBoost classifier. That is, the reported SHAP values show the relative influence of individual features on model predictions, not as evidence of statistical significance or causal relationships. Finally, the statistical significance of the obtained performance differences is analyzed using Friedman and Nemenyi tests, and the findings are discussed.

When the feature groups and the information represented by the original feature set are examined, a total of 75 features belonging to five distinct domains were identified in the dataset (Table 5).

Table 5. Feature groups and their basic characteristics used in the study

Attribute Group	Attribute Number	The Information Represents
Time Domain	21	Basic statistical and amplitude characteristics of the signal
Frequency Domain	24	Spectral distribution and dominant frequency components
Entropy Domain	5	Measures of signal irregularity, uncertainty, and complexity
Nonlinear Domain	5	Signal complexity and dynamic behavior
Wavelet Domain	20	High-resolution time-frequency characteristics
Total	75	Total pre-computed features in the dataset

The RFECV method was used to identify the features that contributed most to classification performance and to reduce the dimensionality of the feature space. The RFECV procedure was applied within the training portion of each outer GroupKFold split. Across the five outer folds, the number of retained features ranged from 41 to 73 features (specifically 47, 73, 55, 47, and 41, with an average of 52.6 features), reflecting fold-specific optimization of the feature space. In each outer fold, only the features selected from the corresponding training partition were used for model training and evaluation on the held-out test partition. The 47-feature configuration was established as the primary baseline for the multi-model benchmark and ablation study (Table 7), whereas the 41-feature configuration reported in Table 10 corresponds to a separate direct-comparison experiment performed using identical GroupKFold partitions. The selected features included time-domain, entropy-domain, frequency-domain, nonlinear, and wavelet-based feature groups, thereby covering different characteristics of the biosensor signals. This approach reduced the computational burden of the model by eliminating unnecessary features while maintaining comparable classification performance. After RFECV, the relative importance levels of the selected features were evaluated using Random Forest feature importance analysis, and the features with the highest contribution are presented in Figure 3.

CC: BY-NC-ND

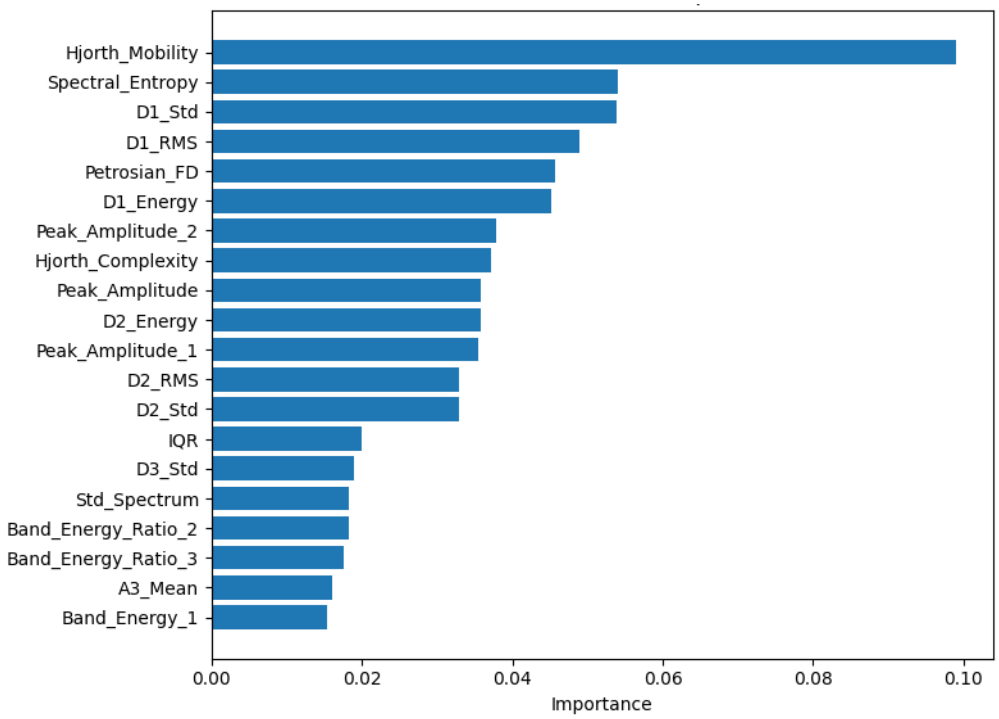


Figure 3. Random Forest feature importance values of the most important features selected after RFECV

When the graph is examined, it shows that the Hjorth_Mobility feature, which represents the rate of change in the frequency content of the signal, contributes significantly more than all other features with an approximate feature attribution of 0.10, indicating that it is the most distinctive feature in the classification process. Hjorth_Mobility is followed by Spectral_Entropy, D1_Std, D1_RMS, Petrosian_FD, and D1_Energy features, respectively. These features contribute to the model at a similar level. The high ranking of spectral and nonlinear complexity measures such as Spectral Entropy and Petrosian Fractal Dimension shows that the irregularity and complex structure of the signal carry important information in the classification success. Similarly, the high feature attributions of wavelet-based statistical features such as standard deviation, RMS, and energy at the D1 level reveal that the high-frequency components of the signal contain distinctive characteristics. Feature importance (Peak_Amplitude_2, Higher_Complexity, Peak_Amplitude, D2_Energy, Peak_Amplitude_1, D2_RMS, and D2_Std) contribute significantly to the model, but to a more limited extent compared to the first group. The selection of energy and RMS-based features, particularly those at the D2 level, indicates that not only high-frequency but also medium-frequency components contain information relevant to classification. Features located at the lower levels of the graph contribute less to the model, but they are not completely eliminated in the RFECV process, providing complementary information. In ensemble learning methods like Random Forest, even feature attribution can contribute to classification performance through interaction with other features.

Hjorth Mobility reflects the relative frequency content of the sensor signal by quantifying the relationship between the variance and the first derivative of the signal [22]. In biosensor measurements, interactions between heavy metal ions and the sensing interface can alter the temporal dynamics of the acquired signal, leading to measurable variations in signal mobility [23]. The high SHAP feature importance of Hjorth Mobility indicates that differences in signal kinetics constitute an informative feature for distinguishing between environmental toxins.

Petrosian fractal dimension quantifies waveform irregularity by measuring sign changes in the first derivative and is widely used to characterize signal complexity[24]. The high SHAP importance assigned to Petrosian fractal dimension suggests that these complexity variations provide discriminative information for the proposed classification model.

CC: BY-NC-ND

Entropy measures quantify the irregularity and complexity of time series and have been widely used to characterize nonlinear dynamics in physiological signals[25]. Because different heavy metal ions interact with the sensing surface through distinct adsorption and binding mechanisms, they may produce sensor responses with different complexity characteristics[26]. The high SHAP feature importance of entropy-based features indicates that signal complexity provides complementary discriminant information in the proposed classification model.

The feature set obtained with RFECV is in total a balanced inclusion of different feature groups (Hjorth parameters, spectral measures, wavelet based statistics, fractal dimension, and peak amplitude measurements). The Hjorth, spectral and wavelet based features are the features with the highest feature attributions, which means that the classification problem is solved better when considering information from both the time and frequency domains together. Additionally, the highest feature attribution is still around 0.10, and other features also contribute significantly to the model, indicating that the model performance does not rely on a single variable, but on the combined effect of multiple and complementary features. This is in line with the fact that RFECV method generates a more compact and generalizable feature set by removing unnecessary features and minimizing information loss.

To eliminate potential feature selection leakage and prevent optimistic performance prediction, the RFECV procedure was performed specifically within the training portion of each outer GroupK fold. In this nested evaluation framework, feature selection was performed using only the training data of each fold, while the corresponding retained test fold remained completely invisible during the selection process.

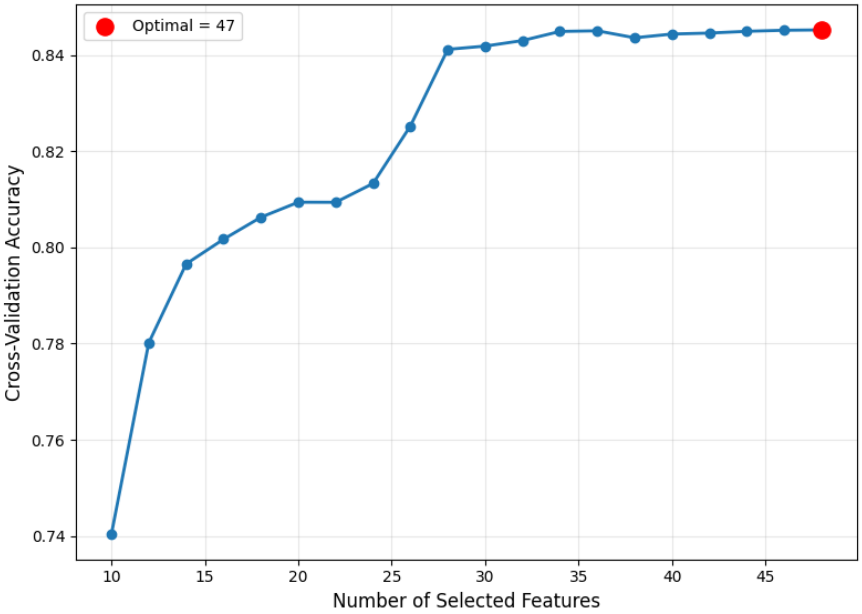


Figure 4. RFECV feature selection optimization curve using nested GroupKFold cross-validation.

Figure 4 illustrates the relationship between the number of features selected and the average cross-validation accuracy obtained during the RFECV procedure. As the number of selected features increased, the classification performance generally increased, reaching a representative plateau around an optimal subset of 47 features. Beyond this point, the addition of further features resulted in only negligible improvements in prediction performance, suggesting that the remaining variables offered limited additional predictive value. Consequently, the number of RFECV-selected features across the outer folds ranged from 41 to 73 features (specifically 47, 73, 55, 47, and 41, with an average of 52.6 features). For each outer fold, the corresponding model was trained and evaluated strictly using only the feature subset selected from that specific fold's training partition.

CC: BY-NC-ND

3.1. Classification results for different models

The performance measures of the six ensemble-based machine learning algorithms after applying the proposed feature selection technique are presented in Table 6 in terms of Accuracy, Precision, Recall, Weighted F1-score, MCC, and Cohen's Kappa. The HistGradientBoosting classifier yielded the best classification accuracy of 85.27%. Unless otherwise stated, Precision, Recall, and Weighted F1-score reported throughout this study correspond to weighted-averaged metrics (average="weighted"), thereby providing an accurate representation across the sample distribution. It also demonstrated slightly better performance than other models in Precision (0.8560), Weighted F1 score (0.8525), MCC (0.8162), and Kappa (0.8153) metrics. The standard deviation of the accuracy is 0.0063, which indicates that the model gives quite stable results across cross-validation folds. Random Forest came in a close second (85.26%), which shows that features selected after RFECV can be effectively used by decision tree-based ensemble methods. LightGBM (85.24%) and XGBoost (85.23%) came next. These four batch learning algorithms showed only small differences in performance, which means that they performed in a stable and competitive way on the biosensor signal dataset. In contrast, the CatBoost model showed the lowest classification accuracy (84.97%) among the models explored. However, the MCC (0.8124) and Kappa (0.8115) values of CatBoost are only about 0.004 points lower than the other models, which shows that the overall classification success of CatBoost is still at a high level. The relatively small performance differences between the models indicate a strong representational capability of the selected feature set and prove that the quality of the features is the main reason for classification success.

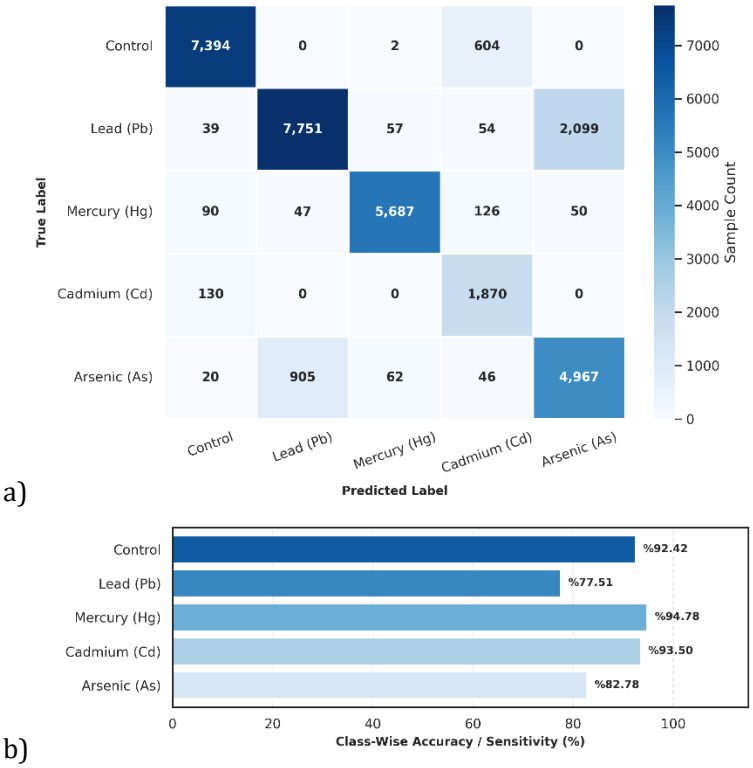
Table 6. Performance metrics of the benchmark machine learning models

Model	Accuracy	Accuracy (SD)	Precision	Recall	Weighted F1	MCC	Kappa
HistGradientBoosting	0.852669	0.006332	0.856024	0.852669	0.852486	0.816160	0.815270
RandomForest	0.852550	0.006308	0.855868	0.852550	0.852305	0.816025	0.815116
LightGBM	0.852431	0.006345	0.855730	0.852431	0.852229	0.815853	0.814969
XGBoost	0.852263	0.006547	0.855498	0.852263	0.852075	0.815621	0.814758
ExtraTrees	0.851169	0.006844	0.854496	0.851169	0.850926	0.814304	0.813393
CatBoost	0.849656	0.006329	0.853208	0.849656	0.849542	0.812410	0.811500

The tested models show quite similar values of precision, recall, Weighted F1-score, MCC and Cohen's Kappa, apart from classification accuracy, which suggests similar predictive performance across multiple evaluation metrics. The standard deviations of the accuracies of all models were between 0.0063 and 0.0068, which indicated that the performance was highly consistent during the cross-validation process and the models were less sensitive to data splitting. Further, the very close precision, recall and Weighted F1-scores show that the models predicted each class in a balanced way and did not exhibit any significant bias towards any specific class. The MCC and Cohen's Kappa coefficients for all models were all greater than 0.81, indicating high accuracy and good fit to random prediction in the classification outcomes. These two metrics are considered as important performance indicators supporting the value of accuracy, especially in balanced datasets.

In addition to prediction performance, computational efficiency was also evaluated by measuring average training and inference times across five GroupKFold folds. HistGradientBoosting required the shortest training time (49.26 s) among all benchmark tree-based ensemble methods, while Random Forest, despite producing comparable prediction accuracy, required a training time of 515.32 s, approximately 12 times longer than HistGradientBoosting. These findings demonstrate that HistGradientBoosting offers a suitable balance between computational efficiency and classification performance.

CC: BY-NC-ND



a)

b)

Figure 5. Quantum toxin classification evaluated via GroupKFold cross-validation a) Confusion matrix b) Per-class recall rates

To rigorously evaluate the multi-class classification performance and reduce potential data leakage, a GroupKFold cross-validation scheme was used. Figure 5 shows the confusion matrix and the corresponding per-class recall rates. The proposed model exhibits robust discrimination capabilities across all target species and achieves particularly high sensitivities for Mercury (94.78%), Cadmium (93.50%), and Control samples (92.42%). Minor inter-class misclassifications were observed mainly between Lead and Arsenic, where 2,099 Lead samples were classified as Arsenic. This overlap is attributed to similar sensor signal decay characteristics under high quantum feature dimensions. These results indicate that the proposed feature representation can capture discriminative signal characteristics across the simulated environmental toxin classes.

In terms of computational efficiency, HistGradientBoosting was much faster to train than Random Forest and gave the highest overall accuracy, showing a favorable tradeoff between predictive performance and computational cost. These results show that HistGradientBoosting offers the best trade-off between classification accuracy and computational efficiency for the proposed biosensor signal classification framework.

We performed ROC curve analysis to evaluate in detail the discrimination performance of the classification models. Figure 6 shows the macro-average ROC curves of six machine learning algorithms for the GroupKFold cross-validation method. The ROC curves are close to the upper left corner, suggesting that the developed models can effectively distinguish the environmental toxin classes.

CC: BY-NC-ND

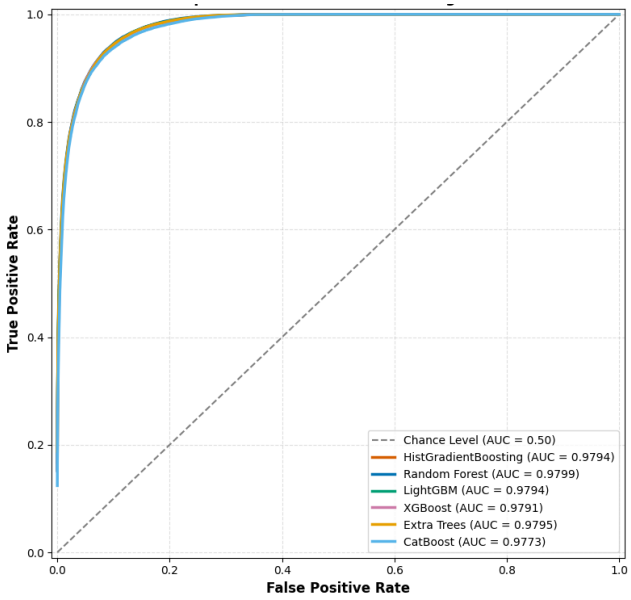


Figure 6. Six machine learning ROC curves and AUC values

ROC analysis showed that Random Forest model had the highest discrimination performance with a macro AUC value of 0.9799. Extra Trees (0.9795), LightGBM (0.9794), HistGradientBoosting (0.9794), XGBoost (0.9791) and CatBoost (0.9773) followed Random Forest. The AUC values of all models were higher than 0.97, which suggests that the proposed RFECV-based feature selection approach has a very strong representation ability.

The algorithms do not show a major difference in performance. The models get high sensitivity (True Positive Rate) and low false positive rate. In particular, the ROC curves of the Random Forest, Extra Trees, LightGBM and HistGradientBoosting models are almost overlapping, showing a similar discrimination performance. However, the CatBoost model, although it has the lowest AUC value of 0.9773, still shows a very successful classification performance.

The ROC analysis results support the accuracy results presented in Table 6. Although the HistGradientBoosting model presented the best performance in terms of accuracy, the Random Forest model showed the best macro AUC value in the ROC analysis, indicating slightly better overall class discrimination across decision thresholds. The difference in AUC between the two models is only 0.0005 which is negligible. This confirms the proposed feature set provides consistent and high performance across different ensemble learning algorithms.

Although the HistGradientBoosting model achieved the highest accuracy for classification, the Random Forest model performed best in distinguishing the classes, with the highest macro AUC value. The LightGBM model, in contrast, showed the most consistent performance across the cross-validation folds, with the least rank value in the average rank analysis. When these findings are considered together, it is seen that the HistGradientBoosting, Random Forest, LightGBM, and XGBoost algorithms show quite similar and competitive performance, and the differences between them may vary depending on the performance metric used. Therefore, it can be said that the proposed feature selection framework approach is not dependent on a specific classifier and can achieve similarly high success rates on different ensemble learning algorithms.

CC: BY-NC-ND

Table 7. Domain-based distribution of the initial 75 pre-computed features and the 47-feature subset identified by RFECV.

Feature Domain	Initial	Selected	Removed	Retention
Time Domain	21	14	7	66.67%
Frequency Domain	24	13	11	54.17%
Entropy Domain	5	3	2	60.00%
Nonlinear Domain	5	4	1	80.00%
Wavelet Domain	20	13	7	65.00%
Total	75	47	28	62.67%

As presented in Table 7, the initial feature set comprised 75 dataset-provided features distributed across five distinct signal domains. The 47-feature baseline was selected by RFECV, retaining complementary attributes from all five domains. The Nonlinear Domain exhibited the highest retention rate (80.00%), followed by the Time Domain (66.67%), Wavelet Domain (65.00%), Entropy Domain (60.00%), and Frequency Domain (54.17%).

3.2. XAI Results

For reliable decision support systems, it is important that machine learning models not only produce predictions with high accuracy but also that the features on which these predictions are based can be explained. Therefore, in this study, SHAP analysis was applied within the scope of the XAI approach. First, the interactions between the basic statistical features were examined using the SHAP interaction graph, and then the average absolute SHAP values were calculated to obtain the global feature importance ranking of the model.

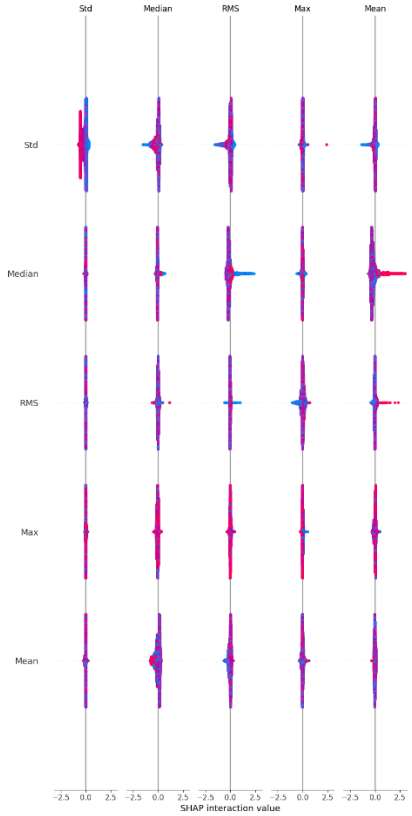


Figure 7. SHAP interaction matrix of the most important features

Figure 7 shows the SHAP interaction values. This graph does not show the ranking of the

CC: BY-NC-ND

importance of individual features, but rather the combined effect of selected basic statistical features (Std, Median, RMS, Max, and Mean) on the prediction. In the SHAP graph, red represents high values of the relevant feature, and blue represents low values. Positive SHAP values increase the model prediction towards the relevant class, while negative values have a reducing effect. When the graph is examined, it is seen that the SHAP interaction values are concentrated around zero in most feature pairs. This indicates that these basic statistical features mostly affect the model decision independently, and there are no strong nonlinear interactions between them. However, it is observed that the interaction distribution is spread over a wider range, especially in the Median–Mean and RMS–Median feature pairs, compared to other combinations. This means that these features might jointly affect the model prediction for certain samples.

The results show that the basic statistical features give complementary information within the model, but the decision-making mechanism is not only based on these variables. Therefore, a global SHAP feature importance analysis was conducted to determine which features contribute decisively to the model performance.

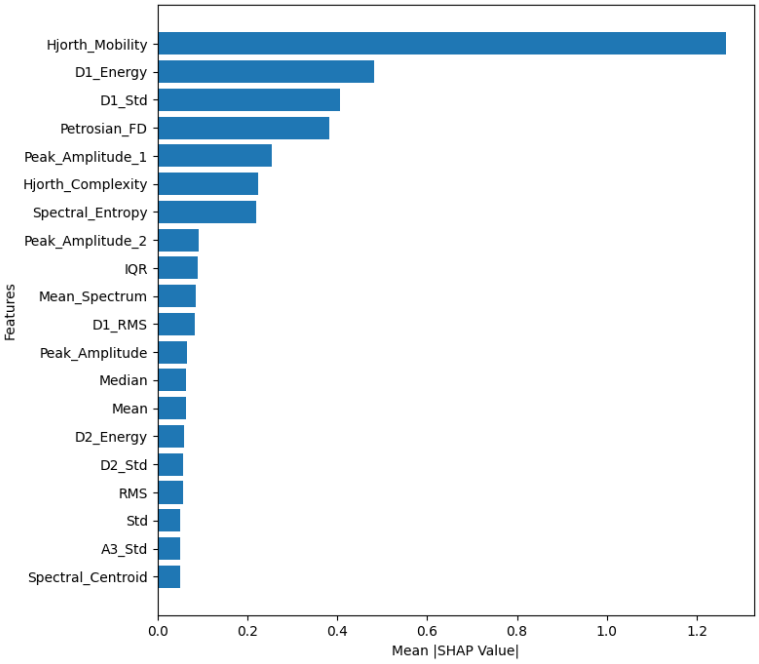


Figure 8. Global feature importance ranking

To improve the interpretability of the proposed framework, SHAP analysis was performed using the TreeExplainer algorithm on the trained XGBoost classifier. SHAP values were calculated for independent test samples (X_test) generated during the evaluation process. Since the classification problem consists of five classes, SHAP values were estimated separately for each class. The global feature importance ranking based on average absolute SHAP values is shown in Figure 8. The average absolute SHAP value for each feature was obtained by averaging the absolute SHAP values across all test samples and all classes. This aggregation provides a model-level interpretation of each feature's overall contribution to the multi-class estimation task. XGBoost was chosen for SHAP analysis because it is a tree-based ensemble model with local and computationally efficient support for TreeExplainer that reliably estimates feature contributions in multi-class classification.

This analysis shows what features the model is making best use of over the whole dataset. From the results, we see that the mean absolute SHAP value of Hjorth_Mobility was substantially higher than for all other variables. This result suggests that the mobility information of biosensor signal is the most potent predictor to discriminate environmental toxin classes. The Hjorth_Mobility feature is followed by D1_Energy, D1_Std, Petrosian_FD, Peak_Amplitude_1, Hjorth_Complexity, and Spectral_Entropy, respectively. Most of these features are based on nonlinear analysis, wavelet transform and frequency

domain analysis. Thus, it appears that the model decision mechanism is determined not only by simple statistical indicators, but also by features describing the dynamic structure and multi-resolution properties of the signal. In contrast, classical statistical features such as Mean, Median, RMS, Std, Spectral_Centroid have lower SHAP feature attribution values. These variables are not sufficient for discrimination by themselves, but complement the model when used in conjunction with other features.

Overall, the SHAP analysis supports the success of the proposed multi-domain feature representation approach. In particular, the combined use of nonlinear signal characteristics, wavelet-based energy components, and spectral features significantly contributes to the model's high classification performance. Furthermore, the SHAP analysis enhances the interpretability of the proposed method by clarifying which features the model's decisions are based on.

3.3. Ablation Study

In the proposed feature evaluation framework, an ablation study was conducted to evaluate the contribution of different feature groups to the classification performance. In this context, the fixed primary 47-feature benchmark baseline ("All Features") was taken as a reference, and then the model was retrained by extracting time-domain, frequency-domain, entropy, nonlinear, and wavelet-based feature groups separately from the dataset. The same HistGradientBoosting classifier and 5-fold GroupKFold cross-validation method were used in all experiments (Figure 9).

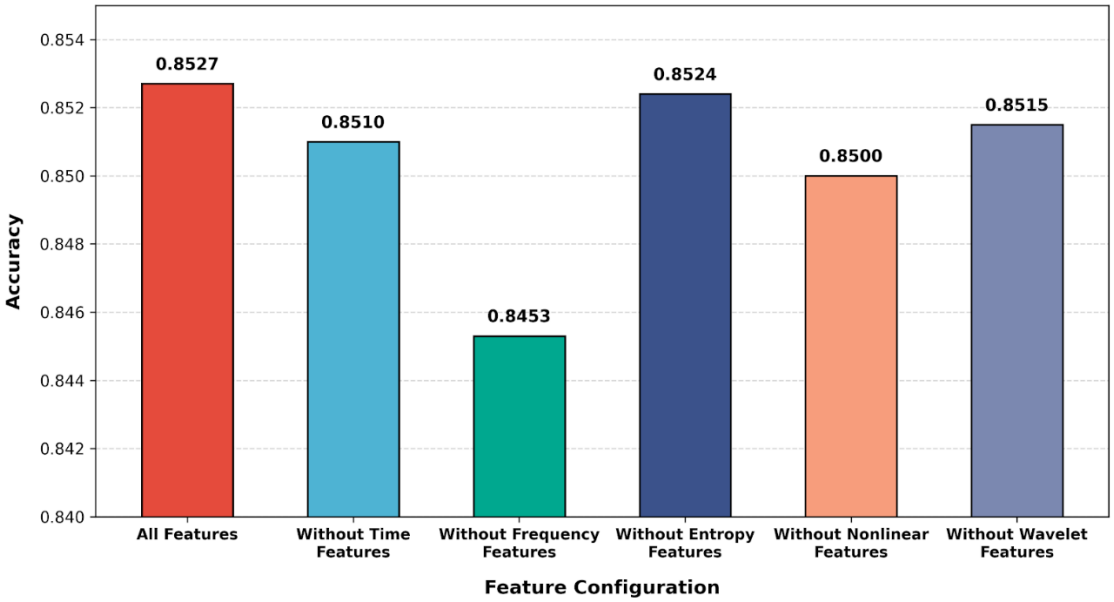


Figure 9. Effect of different feature groups on classification accuracy

Table 8. Classification performance under different feature ablation configurations

Experiment	Feature Count	Accuracy	Accuracy (SD)
All Features	47	0.8527	0.0063
Without Entropy Features	44	0.8524	0.0065
Without Wavelet Features	34	0.8515	0.0065
Without Time Features	33	0.8510	0.0059
Without Nonlinear Features	43	0.8500	0.0056
Without Frequency Features	34	0.8453	0.0058

When the results are examined, the highest accuracy (85.27%) was obtained when the fixed primary 47-feature baseline ("All Features") was used. Removing any feature domain resulted in a

CC: BY-NC-ND

decrease in classification performance. The most substantial performance drop occurred when frequency-domain features were removed, yielding the lowest accuracy of 84.53% and establishing the frequency domain as the primary driver of classification success. Removing nonlinear features resulted in the second largest performance degradation, dropping accuracy to 85.00%. In comparison, removing time-domain and wavelet features resulted in accuracies of 85.10% and 85.15%, respectively, while the smallest performance change occurred after removing entropy features (85.24%). To thoroughly evaluate the contribution of the nonlinear feature group, a Wilcoxon signed-rank test was applied by comparing the fold classification accuracies obtained using the full 47-feature set and the configuration without nonlinear features (85.00% accuracy). Therefore, these results indicate a positive trend and substantial contribution of nonlinear features to model performance within the limits of the 5-fold GroupKFold cross-validation setup.

3.4. Statistical Comparison of Classifiers

Friedman test was applied to determine whether the performances of the six machine learning algorithms used in the proposed method showed statistically significant differences. Accuracy values obtained during the five-fold GroupKFold cross-validation process were used in the analysis. As a result of the Friedman test, the test statistic was obtained as $\chi^2 = 13.5714$ and the significance level was $p = 0.0186$. The obtained p value is less than 0.05, indicating that there are statistically significant performance differences between the classification algorithms examined. Therefore, the null hypothesis that all algorithms perform similarly was rejected.

After the Friedman test, a lower average rank value calculated to evaluate the relative performance of algorithms indicates that the algorithm is more successful[27]. According to the results obtained, LightGBM ranked first with an average rank of 2.4, followed by XGBoost (2.6), HistGradientBoosting (2.8), and Random Forest (2.8) models, respectively. Extra Trees (4.6) and CatBoost (5.8) showed lower performance compared to other models. The average rank results are presented in Figure 10.

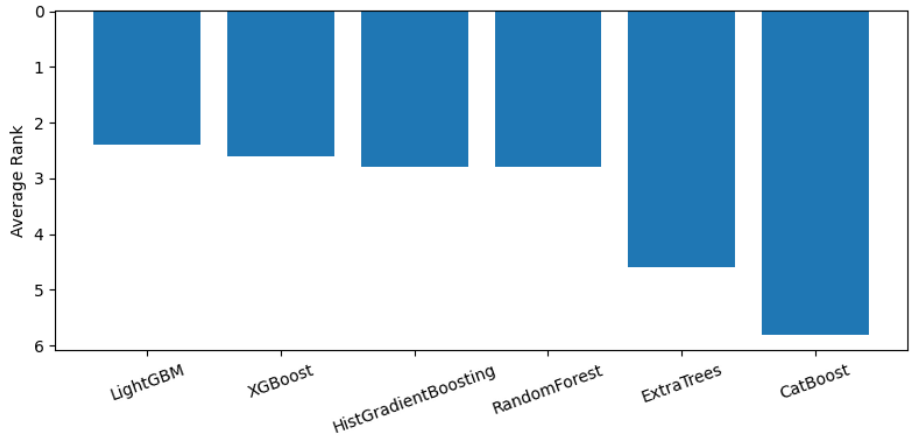


Figure 10. Average rank values of the compared classifiers

Nemenyi post-hoc analysis was used for pairwise comparisons between algorithms. It shows that LightGBM, XGBoost, HistGradientBoosting, and Random Forest models have similar success levels and belong to the same performance group. Although the CatBoost model had the highest numerical average rank (indicating lower relative rank performance), its performance was the lowest. The Extra Trees model performed moderately well and lagged behind the best algorithms. Looking at the Critical Difference diagram, it can be seen that LightGBM, XGBoost, HistGradientBoosting, and Random Forest form a statistically indistinguishable group and exhibit comparable prediction performance. The numerical critical difference (CD) at $\alpha = 0.05$ was approximately 3.37, based on six classifiers and five cross-validation folds. The Nemenyi analysis results are shown in Figure 11 and Table 9. Overall, the Nemenyi results indicate that most classifiers belong to statistically indistinguishable performance groups, despite small numerical differences in their mean performance and average ranks. Therefore,

the observed differences in classification accuracy should not be interpreted as evidence of definitive model superiority unless supported by statistical significance.

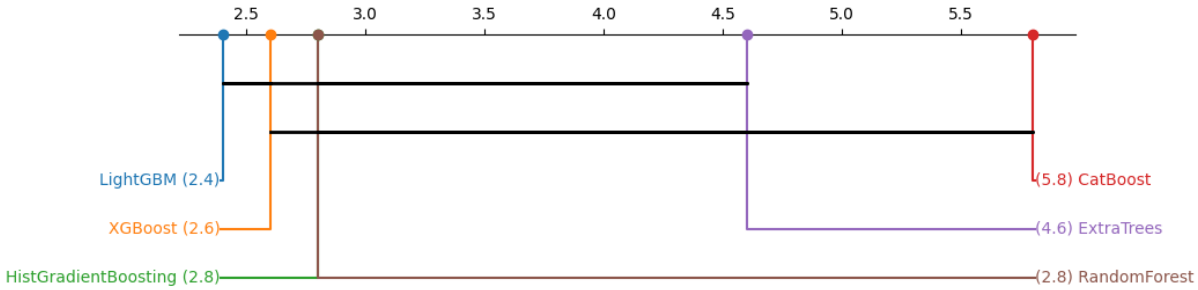


Figure 11. Critical difference diagram of the Nemenyi post-hoc test

Table 9. Pairwise p-values from the Nemenyi post-hoc test

Model	XGBoost	LightGBM	CatBoost	ExtraTrees	RandomForest	HistGradientBoosting
XGBoost	1.000.000	0.999981	0.074302	0.538193	0.999981	0.999981
LightGBM	0.999981	1.000.000	0.046736	0.427525	0.999420	0.999420
CatBoost	0.074302	0.046736	1.000.000	0.913410	0.113891	0.113891
ExtraTrees	0.538193	0.427525	0.913410	1.000.000	0.650490	0.650490
RandomForest	0.999981	0.999420	0.113891	0.650490	1.000.000	1.000.000
HistGradientBoosting	0.999981	0.999420	0.113891	0.650490	1.000.000	1.000.000

All pairwise comparisons except the one between LightGBM and CatBoost yielded p-values greater than 0.05, indicating that the observed performance differences between these model pairs are not statistically significant. Rather than interpreting these models as demonstrating absolute superiority over one another, they should be considered statistically comparable.

Table 10. Direct comparison of the original 75-feature set and the RFECV-selected feature set using identical GroupKFold partitions.

Feature Set	No. Features	Accuracy (Mean ± SD)	Macro-Precision (Mean ± SD)	Macro-Recall (Mean ± SD)	Macro-F1 (Mean ± SD)	MCC (Mean ± SD)	Kappa (Mean ± SD)	Training Time (s)	Inference Time (s)
Original 75 Features	75	0.8406 ± 0.0743	0.8222 ± 0.0732	0.8479 ± 0.0589	0.8290 ± 0.0701	0.7935 ± 0.0963	0.7924 ± 0.0969	6.44 ± 0.98	0.1441 ± 0.0546
RFECV Selected Features	41	0.8373 ± 0.0773	0.8174 ± 0.0764	0.8428 ± 0.0625	0.8236 ± 0.0735	0.7894 ± 0.0998	0.7882 ± 0.1006	4.55 ± 1.06	0.1303 ± 0.0156

In a direct comparison between the original 75-feature set and the RFECV-selected 41-feature subset (Table 10), both configurations were evaluated using the same classifier, identical five-fold GroupKFold partitions, and the same performance metrics. The original 75-feature configuration achieved slightly higher average accuracy than the RFECV-selected configuration. Similar minor differences were observed for Macro-F1 and MCC. Therefore, RFECV did not improve prediction performance in this direct comparison. However, reducing the feature set from 75 to 41 reduced the average training time from 6.44 s to 4.55 s and the average inference time from 0.1441 s to 0.1303 s. These findings demonstrate that RFECV provides a more compact representation with lower computational cost, and its classification performance is close to that of the full feature set.

CC: BY-NC-ND

3.5. Study Limitations and Future Directions

While the proposed framework offers high accuracy and computational efficiency, some methodological limitations must be considered. The sensor signals analyzed in this study were generated using the Qiskit Aer simulator. Although this approach allows for controlled benchmarking, physical quantum biological sensors in real-world environments can encounter dynamic environmental interference, temperature fluctuations, and signal drift that cannot be fully captured by synthetic noise profiles. The current framework was specifically trained on modeled 5-class heavy metal toxin distributions. To verify generalizability across various field environments, validation on experimental wet laboratory biosensor data and hardware-loop quantum processors will be required. While commonly adopted hyperparameter settings were chosen for all benchmark classifiers to ensure a level playing field, extensive hyperparameter optimization was not performed individually for each model. Consequently, the reported benchmark results may not represent the maximum performance achievable by each algorithm. Future work may include Bayesian optimization or other automated hyperparameter search strategies to examine the upper performance limits of the evaluated classifiers in more detail. Furthermore, the statistical power of the rank-based test is limited by the number of cross-validation folds, and number of folds or repeated cross-validation are recommended for future studies.

4. CONCLUSION

This study proposes a comprehensive feature engineering and classification framework for multi-class classification of environmental toxins. A set of 75 pre-calculated multi-domain features provided by the dataset was optimized using RFECV within a Nested GroupKFold-based leak-free cross-validation framework. Subsequently, the fixed 47-feature benchmark configuration was used for benchmarking and ablation analyses. A separate direct comparison was performed between the original 75-feature set and the fold-specific RFECV-selected subset (41 features). When six ensemble learning algorithms were evaluated on the 47-feature baseline, HistGradientBoosting (85.27% accuracy, 0.8525 Weighted F1 score) and Random Forest (85.26%) achieved the highest classification success, with Random Forest providing the highest class separation with a Macro-AUC of 0.9799. SHAP analysis identified Hjorth Mobility, Petrosian Fractal Dimension, D1 Energy, and Spectral Entropy as the features that contributed most to model predictions.

Ablation studies revealed that the most substantial performance degradation occurs when frequency-domain features are eliminated (accuracy dropping from 85.27% to 84.53%), identifying the frequency domain as the primary driver of classification success, followed by nonlinear features (85.00%). Friedman and Nemenyi tests showed that the performance difference between the algorithms was statistically significant, with HistGradientBoosting, Random Forest, LightGBM, and XGBoost models yielding comparable successful results. Overall, the study presents an integrated approach to environmental toxin classification, incorporating a GroupKFold validation strategy to prevent data leakage and SHAP-based explainable AI analysis. This demonstrates that RFECV can provide a significantly more compact feature representation with only a minor change in prediction performance, while reducing computational requirements. Prior to any practical or field application for environmental toxin detection, further experimental validation using physical quantum biological sensors and external wet lab datasets is imperative. This work serves as an initial computational benchmark to guide signal processing architectures for future physical hardware integration.

Peer-review: External, Independent.

Acknowledgements:
The data in the study were taken from the following site.
(https://zenodo.org/records/15345426)

Declarations:

1. Statement of Originality:

This work is original.

2. Author Contributions:

Concept: HİS,KK; ***Conceptualization:*** HİS,KK; ***Literature Search:*** HİS,KK; ***Data Collection:*** HİS,KK; ***Data Processing:*** HİS,KK; ***Analysis:*** HİS,KK; ***Writing – original draft:*** HİS,KK; ***Writing – review & editing:*** HİS,KK.

3. Ethics approval:

Not applicable.

4. Funding/Support:

This work has not received any funding or support.

5. Competing Interests:

The authors declare no competing interests.

6. GenAI Usage Statement:

No GenAI tools were used at any stage of the study.

7. Sustainable Development Goals:

REFERENCES

[1] D. Chakif, J. Furrer, Heavy Metal Toxicity in Clinical and Environmental Health: Sources, Mechanisms, Diagnostics, and Evidence-Based Management of Mercury, Lead, Cadmium, and Arsenic, Int. J. Mol. Sci. 27 (2026). https://doi.org/10.3390/ijms27083513.

[2] K. Keser, M.Ç. Soylu, Detection of Phenylarsine Oxide in Drinking Water Using an Impedimetric Electrochemical Sensor with Gelatin-Based Solid Electrolyte Enriched with Mercaptoethanol: A Novel Prospective Green Biosensor Methodology, ACS Omega 7 (2022) 43111–43121. https://doi.org/10.1021/acsomega.2c05516.

[3] Z. Ling, L. Yang, W. Zhang, T. Yao, H. Xu, Detection of food contaminants: A review of established rapid analytical techniques and their applications and limitations, Food Safety and Health 2 (2024) 72–95. https://doi.org/10.1002/fsh3.12032.

[4] J. Yan, R. Liang, W. Yan, X. Wang, Machine Learning-Enhanced MEC Sensors with Feature Engineering for Quantitative Analysis of Multi-Component Toxicants, Biosensors (Basel). 16 (2026). https://doi.org/10.3390/bios16030144.

[5] S.A. Lahari, N. Kumawat, K. Amreen, R.N. Ponnalagu, S. Goel, IoT integrated and deep learning assisted electrochemical sensor for multiplexed heavy metal sensing in water samples, NPJ Clean Water 8 (2025). https://doi.org/10.1038/s41545-025-00441-x.

[6] M. Kim, S. Adjei-Nimoh, J. Park, S.S. Egon, E. Choi, G. Pak, K. You, A.H.M.A. Sadmani, K.T. Kim, W.H. Lee, Calibration-free on-site detection of microcystin-LR using integrated biosensing, multi-parameter water quality monitoring, and machine learning, Water Res. 298 (2026). https://doi.org/10.1016/j.watres.2026.125832.

[7] K. Tafadzwa Mpofu, P. Mthunzi-Kufa, Recent Advances in Quantum Biosensing Technologies, in: 2024. https://doi.org/10.5772/intechopen.1007858.

[8] U. Zvi, D.R. Candido, A.M. Weiss, A.R. Jones, L. Chen, I. Golovina, X. Yu, S. Wang, D. V. Talapin, M.E. Flatté, A.P. Esser-Kahn, P.C. Maurer, Engineering spin coherence in core-shell diamond nanocrystals, Proc. Natl. Acad. Sci. U. S. A. 122 (2025). https://doi.org/10.1073/pnas.2422542122.

[9] S. Srivastava, W. Wang, W. Zhou, M. Jin, P.J. Vikesland, Machine Learning-Assisted Surface-Enhanced Raman Spectroscopy Detection for Environmental Applications: A Review, Environ. Sci. Technol. 58 (2024) 20830–20848. https://doi.org/10.1021/acs.est.4c06737.

[10] H. Tan, J. Jin, C. Fang, Y. Zhang, B. Chang, X. Zhang, H. Yu, W. Shi, Deep Learning in Environmental Toxicology: Current Progress and Open Challenges, ACS ES&T Water 4 (2024) 805–819. https://doi.org/10.1021/acsestwater.3c00152.

[11] K. Jaganathan, H. Tayara, K.T. Chong, An Explainable Supervised Machine Learning Model for Predicting Respiratory Toxicity of Chemicals Using Optimal Molecular Descriptors, Pharmaceutics 14 (2022). https://doi.org/10.3390/pharmaceutics14040832.

[12] C.-C. Chang, Y. Chen, 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. 297 (2026) 124207. https://doi.org/10.1016/j.envres.2026.124207.

[13] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

[14] M. Khajavian, J.-H. Jang, J.-Y. Kwon, J.-M. Lee, S. Lee, M.-H. Hwang, E. Yang, J.K. Jang, K.-J. Chae, Harnessing interpretable machine learning: SHapley additive exPlanations (SHAP)-driven insights, transformative impact, and controversies in adsorption-based environmental remediation, Inorg. Chem. Commun. 186 (2026) 116269. https://doi.org/10.1016/j.inoche.2026.116269.

[15] L.T. Yogarathinam, S.I. Abba, J. Usman, D.U. Lawal, I.H. Aljundi, Predicting micropollutant removal through nanopore-sized membranes using several machine-learning approaches based on feature engineering, RSC Adv. 14 (2024) 19331–19348. https://doi.org/10.1039/d4ra02475c.

[16] https://zenodo.org/records/15345426, (n.d.).

[17] Pagidipalla, M. S. (2025). Dataset and Code for Quantum-AI Enhanced Biosensor for Environmental Toxin Detection: A Synthetic Data Approach [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.15345426, (n.d.).

[18] 10.5281/zenodo.15345425, (n.d.).

[19] M. Awad, S. Fraihat, Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems, Journal of Sensor and Actuator Networks 12 (2023). https://doi.org/10.3390/jsan12050067.

[20] C. Chen, J. Liang, W. Sun, G. Yang, X. Meng, An automatically recursive feature elimination method based on threshold decision in random forest classification, Geo-Spatial Information Science 28 (2025) 1494–1519. https://doi.org/10.1080/10095020.2024.2387457.

[21] Z. Zheng, Y. Jiang, Q. Zhang, Y. Zhong, L. Wang, A Feature Selection Method Based on Relief Feature Ranking with Recursive Feature Elimination for the Inversion of Urban River Water

CC: BY-NC-ND

Quality Parameters Using Multispectral Imagery from an Unmanned Aerial Vehicle, Water (Switzerland) 16 (2024). https://doi.org/10.3390/w16071029.

[22] W.H. Alawee, A. Basem, L.A. Al-Haddad, Advancing biomedical engineering: Leveraging Hjorth features for electroencephalography signal analysis, J. Electr. Bioimpedance 14 (2023) 66–72. https://doi.org/10.2478/joeb-2023-0009.

[23] G.L. Turdean, Design and Development of Biosensors for the Detection of Heavy Metal Toxicity, International Journal of Electrochemistry 2011 (2011) 1–15. https://doi.org/10.4061/2011/343125.

[24] A.F. Mohamed, V. Jusas, Advancing Fractal Dimension Techniques to Enhance Motor Imagery Tasks Using EEG for Brain–Computer Interface Applications, Applied Sciences (Switzerland) 15 (2025). https://doi.org/10.3390/app15116021.

[25] Y. Ma, W. Shi, C.K. Peng, A.C. Yang, Nonlinear dynamical analysis of sleep electroencephalography using fractal and entropy approaches, Sleep Med. Rev. 37 (2018) 85–93. https://doi.org/10.1016/j.smrv.2017.01.003.

[26] C. Fan, W. Li, J. Ning, D. Yang, Z. Zhu, W. Cheng, Y. Liang, H. Hu, Interfacing chemistry and materials for heavy metal ion sensing: mechanistic foundations and adaptive design strategies, Mater. Chem. Front. 10 (2026) 1368–1408. https://doi.org/10.1039/D6QM00046K.

[27] J. Demšar, Statistical Comparisons of Classifiers over Multiple Data Sets, 2006.

CC: BY-NC-ND