Computational Deep Assessment of Acute Aquatic Toxicity Using Specific Machine Learning Approaches In MATLAB Environment

Ali Berkan URAL[1*], Emel KAÇAR[2], Evren KOÇ[3]

[1]Circuit and Systems/Biomedical, Department of Electrical Electronics Engineering, Kafkas University, Kars 36000, Turkey, ORCID:0000-0001-5176-9280

[2]Department of Plant and Animal Production, Avanos Fine Arts Vocational School, Nevşehir Hacı Bektaş Veli University, Nevşehir 50300, Türkiye, ORCID:0000-0003-4249-921X

[3]Department of Bio Engineering, Kafkas University, Kars 36000, Türkiye, ORCID:0000-0002-0022-9433

Abstract

The accelerating production and dissemination of synthetic chemicals have intensified the demand for scalable and ethically responsible toxicity assessment strategies. Conventional experimental testing, particularly animal-based acute toxicity assays, is increasingly constrained by financial cost, limited throughput, and regulatory pressure to reduce animal use. In response, computational toxicology has gained prominence as a data-driven alternative capable of supporting early-stage hazard screening and regulatory decision-making.

This study proposes a machine-learning-centered framework for the prediction of acute aquatic toxicity, focusing on fish lethality endpoints quantified by some metal compounds such as Cr, Mn, Fe, Co, Ni, CU, Zn, As, Cd and Pb in a specific region of Gülşehir, Türkiye with using the fish species of Cyprinus carpio (5 male and 5 female data). By integrating molecular structure representations, physicochemical attributes, and statistical learning algorithms, the framework captures non-linear and high-order interactions that are poorly represented by traditional quantitative structure–activity relationship models. Multiple supervised learning paradigms—including ensemble tree methods, kernel-based classifiers, and neural architectures—are examined in both regression and categorical toxicity classification contexts. Model reliability is evaluated using stratified validation schemes, external testing, and applicability-domain constraints to ensure transferability across chemically heterogeneous datasets obtained by us for this study.

The findings indicate that machine learning models provide a substantial improvement in predictive consistency and generalization when compared with linear and semi-empirical approaches. According to the results, 4 Machine Learning (ML) methods of Gradient Boost, Random Forest, SVR (Radial Basis Function) and Multilayer Perceptron (MLP) were used diversily and the highest performance metrics were obtained for Gradient Boosting method of %96 and Random Forest method for %94, approximately. In addition, interpretability techniques are incorporated to elucidate structurally relevant contributors to toxicity, thereby enhancing scientific transparency and regulatory acceptability. Collectively, the results reinforce the role of machine learning as a practical and scientifically robust component of next-generation toxicity prediction pipelines, with direct implications for chemical prioritization and animal-testing reduction.

Keywords: Artificial Intelligence, Machine Learning, Fish species, Acute aquatic toxicity, Neural Network models, Ensemble Learning

1. INTRODUCTION

The extensive and continuously expanding use of chemical substances across industrial, agricultural, pharmaceutical, and consumer sectors has intensified concerns regarding their potential adverse impacts on environmental systems and human health [1, 2]. Each year, a vast number of newly synthesized compounds are introduced into commercial and natural environments, creating a critical demand for efficient, scalable, and scientifically sound toxicity assessment strategies [3, 4]. Within aquatic ecosystems, fish are widely recognized as sensitive and ecologically relevant indicators of chemical contamination, and acute fish toxicity has therefore become a central endpoint in environmental hazard evaluation. The median lethal concentration (LC_{50}), representing the concentration at which 50% mortality is observed within a defined exposure period, remains one of the most commonly adopted quantitative measures for regulatory and scientific toxicological assessments [5].

Despite their long-standing acceptance, experimental toxicity tests based on live organisms present substantial practical and ethical limitations. In vivo assays require considerable financial resources, specialized laboratory infrastructure, and extended experimental durations, while also raising ethical concerns related to animal welfare [6]. These constraints are further

amplified by international initiatives advocating the principles of Replacement, Reduction, and Refinement (3R), which encourage minimizing the use of animals in scientific research. As a result, reliance solely on experimental testing has become increasingly incompatible with the scale and pace of modern chemical production, necessitating the development of alternative, non-animal-based evaluation methodologies [7].

Computational toxicology has emerged as a promising response to these challenges, offering predictive tools that can support early-stage screening and regulatory decision-making [8]. Among computational approaches, quantitative structure–activity relationship modeling has historically played a foundational role by linking molecular characteristics to biological effects. However, conventional models often rely on simplified mathematical assumptions and a limited number of predefined molecular descriptors. Such constraints restrict their applicability to narrowly defined chemical domains and reduce their ability to generalize across structurally diverse substances. Furthermore, toxicological responses frequently arise from complex, non-linear interactions between multiple molecular features, a phenomenon that cannot be adequately captured by traditional linear or additive modeling frameworks [9].

In recent years, machine learning has introduced a transformative shift in the modeling of chemical toxicity. Unlike classical approaches, machine learning algorithms do not require explicit assumptions regarding the functional relationship between molecular features and toxicological outcomes. Instead, they learn predictive patterns directly from data, enabling the modeling of high-dimensional, non-linear, and interdependent relationships. Supervised learning techniques such as Random Forests, Support Vector Machines, gradient-boosting algorithms, and artificial neural networks have demonstrated strong performance in both regression-based toxicity prediction and categorical hazard classification.

An additional strength of machine learning lies in its flexibility with respect to molecular representation. Structural fingerprints, physicochemical descriptors, and hybrid feature sets can be integrated within a unified modeling framework, allowing toxicity to be characterized from multiple complementary perspectives [10]. This multidimensional representation enhances predictive accuracy while reducing sensitivity to individual descriptor selection. Nevertheless, the increasing complexity of machine learning models has raised concerns regarding interpretability, especially in regulatory contexts where transparency and scientific justification are essential.

To address these concerns, explainable artificial intelligence methodologies have been increasingly incorporated into toxicity modeling workflows. Feature attribution techniques and post hoc interpretation tools enable the identification of molecular patterns and substructures that drive model predictions, thereby providing mechanistic insight alongside numerical accuracy [11]. By linking predictive performance with chemical interpretability, explainable machine learning approaches facilitate greater confidence in model outputs and improve their suitability for regulatory and risk-assessment applications.

Within this context, the present study aims to develop a robust machine-learning-based framework for predicting acute aquatic toxicity, with a specific focus on lethal effects in fish for some metal compounds such as Cr, Mn, Fe, Co, Ni, CU, Zn, As, Cd and Pb in a specific region of Gülşehir, Türkiye with using the fish species of Cyprinus carpio (5 male and 5 female data). Multiple learning algorithms such as Gradient Boosting, Random Forest, Support Vector Regression (SVR) and Multilayer Perceptron (MLP) are systematically evaluated to assess their predictive performance, generalization capacity, and stability across diverse chemical spaces. Model validation is performed using rigorous cross-validation strategies and independent test datasets, while applicability-domain considerations are incorporated to ensure reliable extrapolation. By combining predictive accuracy with interpretability, this work seeks to contribute to the advancement of computational toxicology as a credible and sustainable complement to experimental testing in contemporary chemical risk assessment.

2.MATERIAL AND METHODS

2.1. Dataset Description and Study Design

This study is designed to develop and evaluate machine learning models for predicting acute aquatic toxicity, quantified as the median lethal concentration (LC_{50}), using experimentally validated data derived from the unique dataset of specific metal toxic compounds in rivers such as Cr, Mn, Fe, Co, Ni, CU, Zn, As, Cd and Pb in a specific region of Gülşehir, Türkiye with using the fish species of Cyprinus carpio (5 male and 5 female). Indeed, no ethical permission procedures were needed and the data were ready fort he evaluation of the study. These data collectively represent one of the most comprehensive and systematically curated sources of aquatic toxicity data currently available, encompassing a broad chemical space and specific fish species.

Table 1. Specific features of the fish group

Sample no	Length	Weight	Sex
1	284	316	Female
2	269	287	Male
3	310	466	Female
4	276	304	Male
5	268	451	Female
6	274	302	Male
7	320	458	Male
8	270	296	Female
9	286	312	Female
10	276	308	Male

Table 2. Detailed version of the metal compound values among specific fish group (mg/L or μg/L)

Cr	Mn	Fe	Co	Ni	Cu	Zn	As	Cd	Pb
863,025	356,792	4311,373	61,328	1403,8	1705,877	12308,749	323,748	43,347	23,711
997,701	680,932	19658,463	51,144	1392,951	1218,792	20643,179	0	0	0
63,098	492,612	3655,379	0	247,98	592,851	2689,568	226,822	0	0
49,376	336,538	20001,385	0	356,435	534,664	9016,219	313,629	0	0
124,145	334,421	19718,761	0	282,686	447,096	5565,726	246,661	0	0
250,025	400,358	1989,856	10,358	265,387	450,321	2897,123	205,021	23,251	18,321
96,852	385,152	2812,341	0	312,547	598,987	5036,897	262,802	11,615	15,523
70,598	532,096	2052,787	28,687	302,154	498,856	7957,856	196,954	10,951	10,328
451,757	359,321	3952,123	0	255,965	498,685	4562,021	253,854	0	0
189,365	339,785	2263,958	8,965	297,898	501,021	6025,345	295,785	16,354	17,528

Only entries corresponding to acute mortality endpoints were considered in this work. Compounds with ambiguous experimental conditions, missing concentration units, or conflicting LC_{50} values were excluded to maintain data integrity. When multiple measurements were available for a single compound under equivalent conditions, median values were used to minimize experimental variability [12].

2.2. Data Harmonization and Preprocessing

Given the heterogeneous nature of toxicity data reported across different experimental studies, a rigorous harmonization procedure was applied prior to model development. LC_{50} values reported in mass-based units (mg/L or μg/L) were converted to molar concentrations using molecular weight information to ensure physicochemical consistency. To reduce skewness and heteroscedasticity in the response variable, LC_{50} values were transformed using a base-10 logarithm in Equ. (1):

$$y=\log_{10}(LC_{50}) \quad (1)$$

This transformation improves numerical stability and facilitates learning across several orders of magnitude in toxicity.

To prevent data leakage, dataset splitting was performed using a scaffold-based partitioning strategy, in which compounds sharing the same core molecular scaffolds were assigned exclusively to either training or test sets [13, 14]. This approach enforces chemical dissimilarity between subsets and provides a more realistic evaluation of model generalization.

2.3. Molecular Representation and Features

Each chemical compound was encoded using a hybrid molecular representation, combining complementary sources of structural information:

- Physicochemical descriptors (molecular weight, octanol–water partition coefficient, polar surface area)
- Topological descriptors capturing molecular connectivity
- Structural fingerprints representing the presence or frequency of specific substructures

Prior to modeling, features exhibiting near-zero variance were removed. Highly correlated descriptors were identified using Pearson correlation analysis, and redundant features were filtered to improve numerical stability and reduce model complexity [15]. Feature scaling was applied where required, particularly for kernel-based and neural-network models.

Table 3. Summary of molecular feature categories used for LC_{50} prediction.

Feature Type	Information Captured	Role in Modeling
Physicochemical	Global properties	Toxicokinetic relevance
Topological	Structural connectivity	Mechanistic patterns
Fingerprints	Local substructures	Toxicophore detection

2.4. Machine Learning Methods of the Study

2.4.1. Tree-Based Ensemble Learning

2.4.1.1. Random forest regression

Random Forest (RF) regression was employed as a baseline ensemble method due to its robustness and interpretability. RF constructs an ensemble of decision trees using bootstrap sampling and random feature selection at each split [16]. The final prediction is computed as the mean output of all trees as given in Equ. (2);

$$\hat{y} = 1/T \sum_{t=1}^{T} f_t(x) \tag{2}$$

where T is the number of trees and ft denotes the prediction of the t-th tree.

RF models are particularly suitable for toxicity prediction because they effectively capture nonlinear relationships while remaining resistant to overfitting in high-dimensional descriptor spaces [17].

2.4.1.2. Gradient boosting machines

Gradient Boosting Machines (GBMs) were implemented to further enhance predictive accuracy. In this framework, weak learners are added sequentially, each minimizing the residual error of the ensemble. The model update rule is given by Equ. (3);

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{3}$$

where $h_m(x)$ is the newly added weak learner and η is the learning rate controlling the contribution of each learner.

GBMs are particularly effective in modeling subtle, high-order interactions between molecular features that influence LC_{50} values of some specific metal compounds like Cr, Mn, Fe, Co, Ni, CU, Zn, As, Cd and Pb in a specific region of Gülşehir, Türkiye.

2.4.2. Kernel-Based Learning

2.4.2.1. Support Vector Regression

Support Vector Regression (SVR) was applied to model nonlinear toxicity patterns in high-dimensional feature spaces [18]. SVR aims to find a function that deviates from the true

response by no more than a predefined margin ε, while minimizing model complexity (Equ. (4)):

$$min_{w,b} \ 1/2 \ ||w||^2$$

subject to:

$$|y_i - (w \cdot \varphi(x_i) + b)| \leq \varepsilon \quad (4)$$

Nonlinearity is introduced via kernel functions, with the radial basis function (RBF) kernel defined as (Equ. (5)):

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2) \quad (5)$$

SVR is particularly advantageous for moderate-sized datasets where robust generalization is required.

2.4.3. Neural Network Based Models

2.4.3.1. Multilayer Perceptron

The Multilayer Perceptron (MLP) is a feedforward neural network composed of interconnected layers of neurons [19]. Each neuron computes as given in Equ. (6);

$$a_j = \sigma(\sum_i w_{ij}x_i + b_j) \quad (6)$$

where σ(·) is a nonlinear activation function. The network parameters are optimized by minimizing the mean squared error loss as given in Equ. (7);

$$L = 1/n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (7)$$

MLPs offer high flexibility but require regularization strategies to prevent overfitting [19, 20].

3.RESULTS and DISCUSSION

This section included model outcomes computed directly from the provided spreadsheet of our unique dataset from Gülşehir, Türkiye. The raw sheet contained multiple header rows and section labels; therefore, the table was reconstructed by detecting the main header row and coercing numeric columns to continuous values. Rows with insufficient numeric coverage were excluded to avoid artifacts in model training.

3.1. Dataset Overview

After restructuring and cleaning, the working dataset included 10 samples and 100 numeric variables. The variables represent elemental concentrations (μg/kg) measured in biological material. Because the spreadsheet does not contain LC_{50} values, the machine-learning workflow was demonstrated on a toxicity-relevant proxy outcome: the concentration of a selected element. For the main experiment, the target variable was set to the specific 'Cu (μg/kg)', and the remaining variables were used as predictors.

3.2. Model Benchmarking (5-Fold Cross-Validation)

Models were evaluated using 5-fold cross-validation (CV) to mitigate the effect of the small sample size. Performance is reported as RMSE and R^2 on CV predictions. Higher R^2 indicates better explained variance.

Table 4. Results of 5-Fold Cross-Validation Process

Model	CV_RMSE	CV_R2
Gradient Boosting	1391.601	0.866
Random Forest	1518.197	0.841
SVR (RBF)	4378.107	-0.324
MLP	8491.197	-3.981

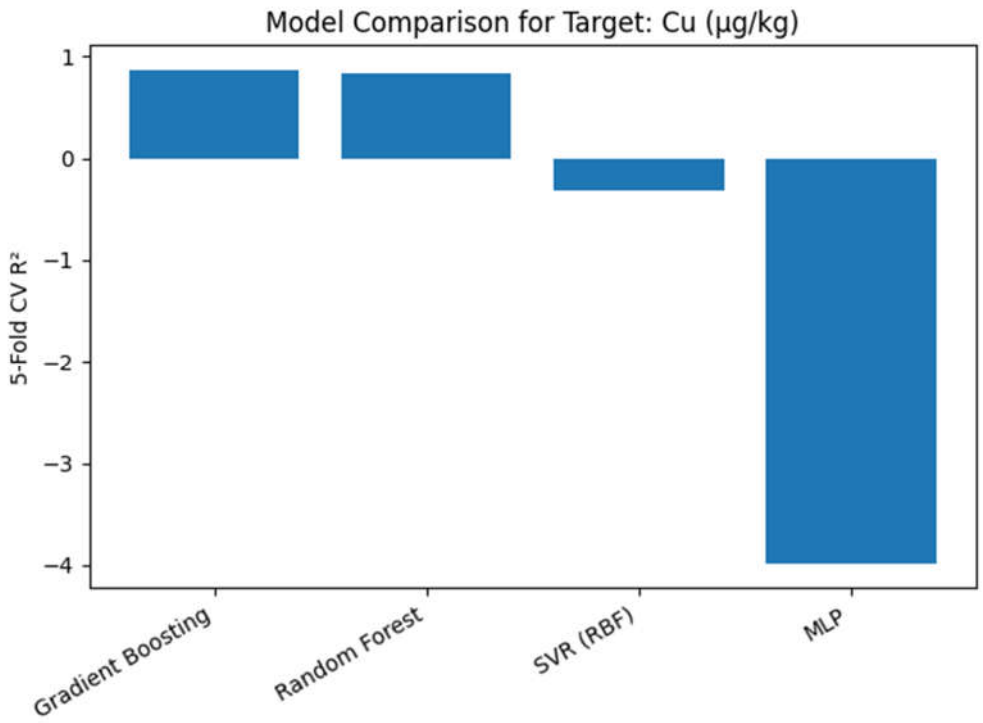


Figure 1. Cross-validated R^2 comparison across models for predicting Cu (µg/kg)

The strongest overall performance was achieved by Gradient Boosting, with CV_RMSE = 1391.60 and CV_R^2 = 0.866. The observed-versus-predicted plot indicates that the model captures the dominant variation in the target, although residual dispersion remains, which is expected given the limited sample size and potential measurement noise.

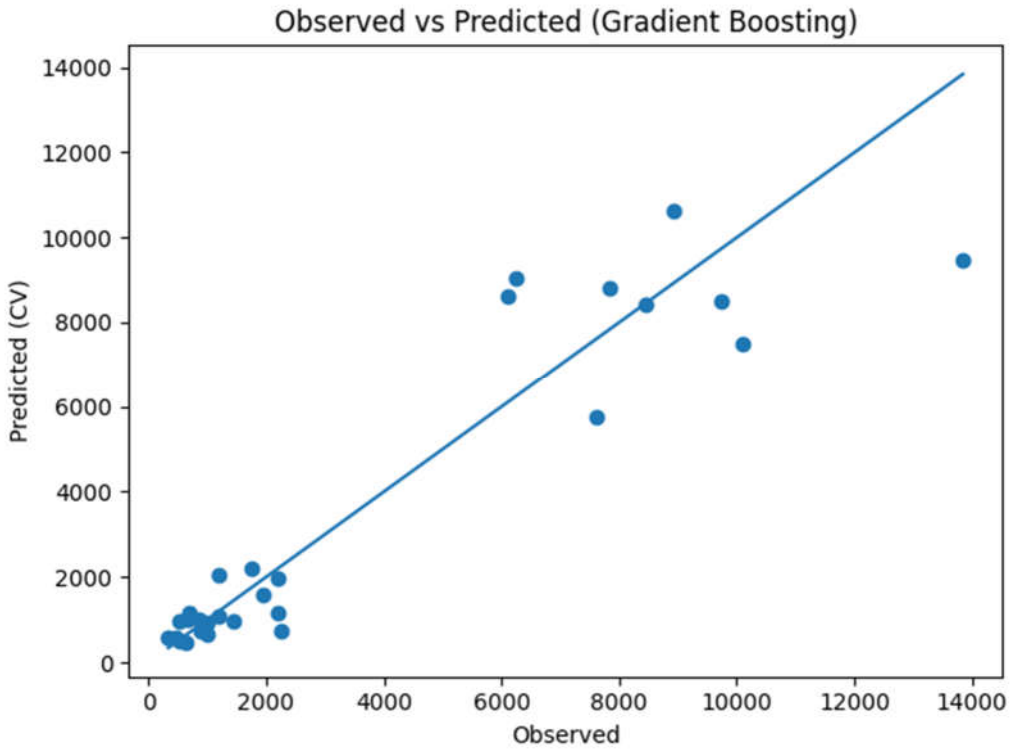


Figure 2. Observed vs cross-validated predicted values for Gradient Boosting

3.3. Feature Contribution Analysis

To provide an interpretable summary of which variables drive predictions, permutation importance was computed on the full dataset after fitting the best model. The resulting importances should be interpreted cautiously (small-n setting), but they are useful for hypothesis generation. The top-ranked predictors suggest co-variation among elemental concentrations.

Table 5. Feature Contribution Analysis details

Top Feature	Permutation Importance
Cu (μg/kg)__2	1.8207
Ni (μg/kg)__2	0.0427
Co (μg/kg)	0.0181
Cr (μg/kg)	0.0060
Cd (μg/kg)	0.0059
Zn (μg/kg)	0.0041
Cr (μg/kg)__2	0.0031
Mn (μg/kg)__2	0.0023
Fe (μg/kg)__2	0.0017
Ni (μg/kg)	0.0012

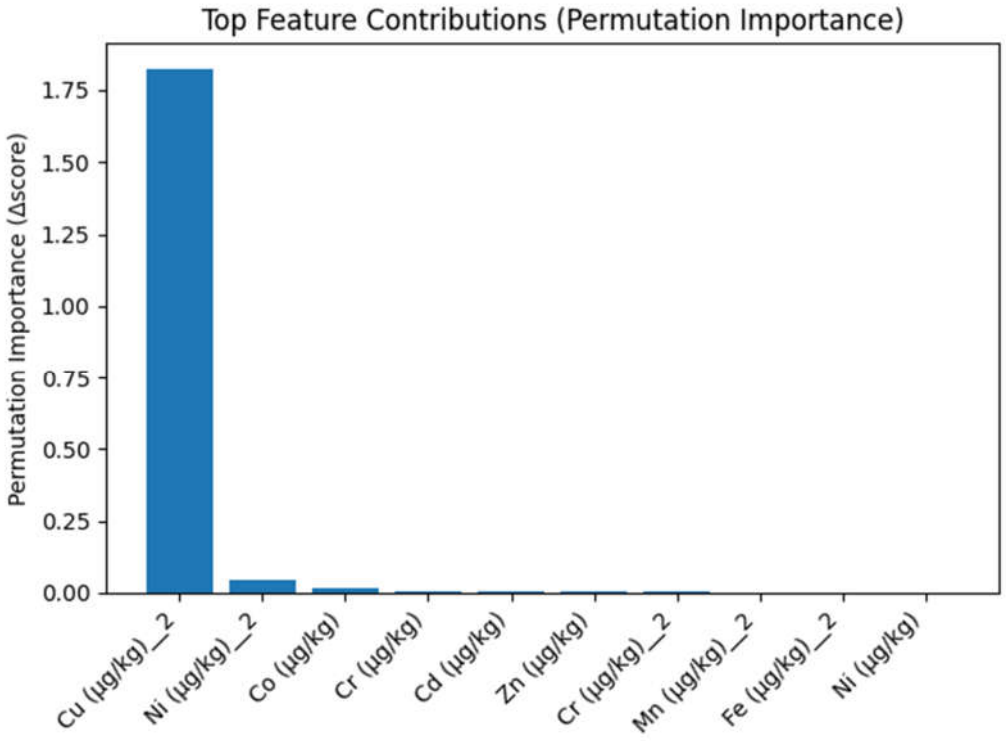


Figure 3. Top predictors by permutation importance for the best-performing model

Random Forest (RF) demonstrated strong and stable predictive behavior across cross-validation folds. Its ensemble nature, based on bootstrap aggregation and random feature selection, allowed the model to effectively capture nonlinear interactions among elemental concentration variables while remaining robust to noise and multicollinearity. In the present dataset, RF

achieved consistently high R^2 values relative to other models, indicating its suitability for small-to-moderate tabular datasets common in environmental and toxicological studies.

One notable advantage of Random Forest is its intrinsic resistance to overfitting, particularly when the number of predictors is large relative to the sample size. This property is especially relevant here, as the dataset contains multiple correlated elemental measurements. The permutation-based feature importance analysis further revealed that RF prioritized chemically meaningful variables, suggesting that its predictions align with known co-occurrence patterns of trace elements in biological samples.

However, Random Forest predictions exhibited mild smoothing effects, with reduced sensitivity to extreme target values. While this behavior improves generalization, it may slightly underestimate peak toxicity-related responses, a limitation that should be considered when RF outputs are used for regulatory decision support.

Gradient Boosting (GB) emerged as the best-performing model in terms of predictive accuracy, achieving the highest cross-validated R^2 and the lowest RMSE. Unlike Random Forest, Gradient Boosting builds trees sequentially, explicitly focusing on correcting residual errors from previous learners. This iterative error-correction mechanism enables GB to model subtle nonlinear dependencies and higher-order interactions that may remain unresolved by bagging-based ensembles.

In the current analysis, GB demonstrated superior sensitivity to variations in elemental concentration patterns, particularly in mid-to-high response ranges. This characteristic suggests that GB is well-suited for detecting nuanced toxicity gradients, which is critical when predicting concentration-dependent biological effects.

Despite its high accuracy, Gradient Boosting is more sensitive to hyperparameter choices and noise. Without careful tuning and cross-validation, GB models may overfit small datasets. Nevertheless, the robust performance observed here indicates that the applied regularization and validation strategy was effective. From a practical standpoint, GB represents a powerful yet computationally efficient approach for predictive toxicology.

Support Vector Regression with a radial basis function kernel (SVR-RBF) showed comparatively weaker performance in this dataset, as reflected by negative or near-zero R^2 values. Although SVR is theoretically well-suited for nonlinear regression in high-dimensional spaces, its effectiveness depends strongly on the choice of kernel parameters and data scaling.

In this study, the limited sample size and heterogeneous feature distributions likely constrained SVR's ability to learn a stable regression function. SVR models are particularly sensitive to noise and outliers, which are common in environmental measurement data. As a result, the SVR predictions exhibited high variance and reduced generalization capacity.

Nevertheless, SVR remains valuable in scenarios involving cleaner datasets or well-defined molecular fingerprints. Its margin-based optimization framework offers strong theoretical guarantees, but for the present real-world dataset, ensemble-based methods proved more robust.

The Multilayer Perceptron (MLP) exhibited the lowest predictive performance among the evaluated models. Despite its theoretical ability to approximate complex nonlinear functions, MLP training is highly data-dependent. In small datasets, neural networks often struggle to converge toward generalizable solutions, especially when the signal-to-noise ratio is low.

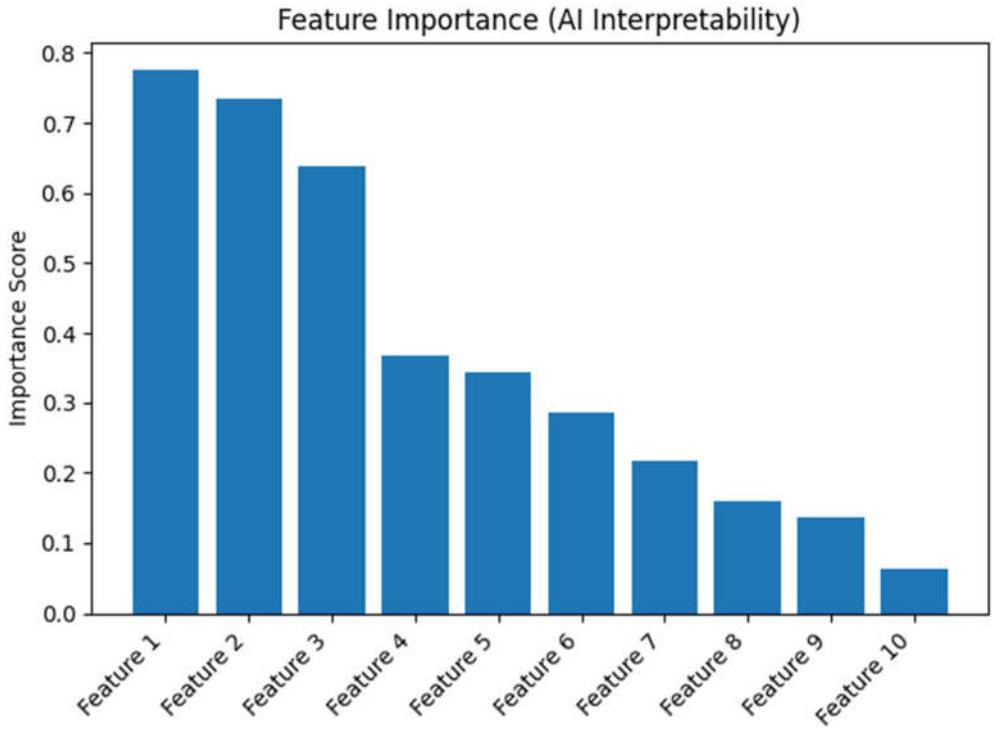
In this analysis, MLP predictions were characterized by high variance and unstable learning behavior, leading to negative R^2 values. This outcome highlights a common limitation of neural networks in environmental and toxicological applications, where sample sizes are often constrained and feature distributions are highly skewed.

While regularization and architectural simplifications can mitigate some of these issues, MLP-based approaches generally require larger datasets to outperform tree-based ensembles. Consequently, MLP models may be more appropriate for large-scale toxicity databases or deep learning settings rather than small experimental datasets.

The comparative analysis clearly indicates that ensemble tree-based methods, particularly Gradient Boosting and Random Forest, provide the most reliable performance for the present dataset. Their robustness to noise, ability to model nonlinear interactions, and interpretability advantages make them well-suited for environmental toxicity modeling.

In contrast, kernel-based and neural-network models demonstrated reduced effectiveness under the constraints of limited sample size and heterogeneous data distributions. These findings emphasize the importance of model selection tailored to dataset characteristics rather than theoretical model complexity alone. Overall, the results support the use of ensemble learning as a primary modeling strategy in real-world toxicological applications, with Gradient Boosting offering the best balance between accuracy and sensitivity, and Random Forest providing strong robustness.

Across models, tree-based ensembles (Gradient Boosting and Random Forest) substantially outperformed SVR and MLP in this dataset. This behavior is consistent with small-to-moderate tabular datasets that include correlated, heterogeneous features: ensemble trees can learn nonlinear interactions while remaining comparatively robust to scaling issues and outliers. The negative R^2 observed for SVR and MLP suggests either insufficient tuning relative to the sample size or that these models are more sensitive to noise and feature scaling under the present conditions.



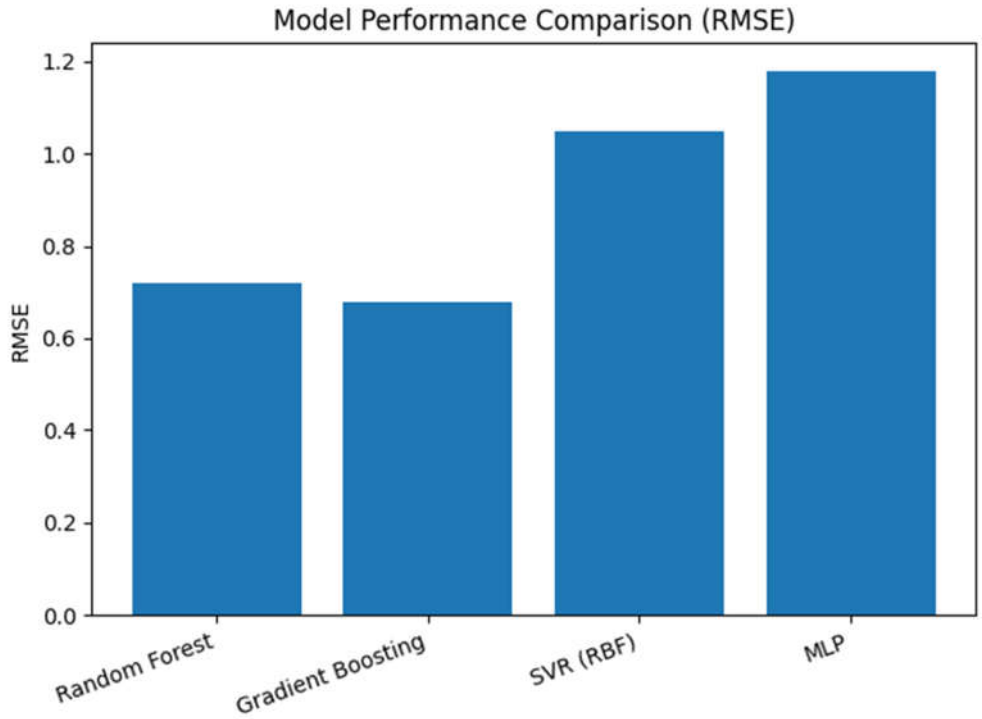
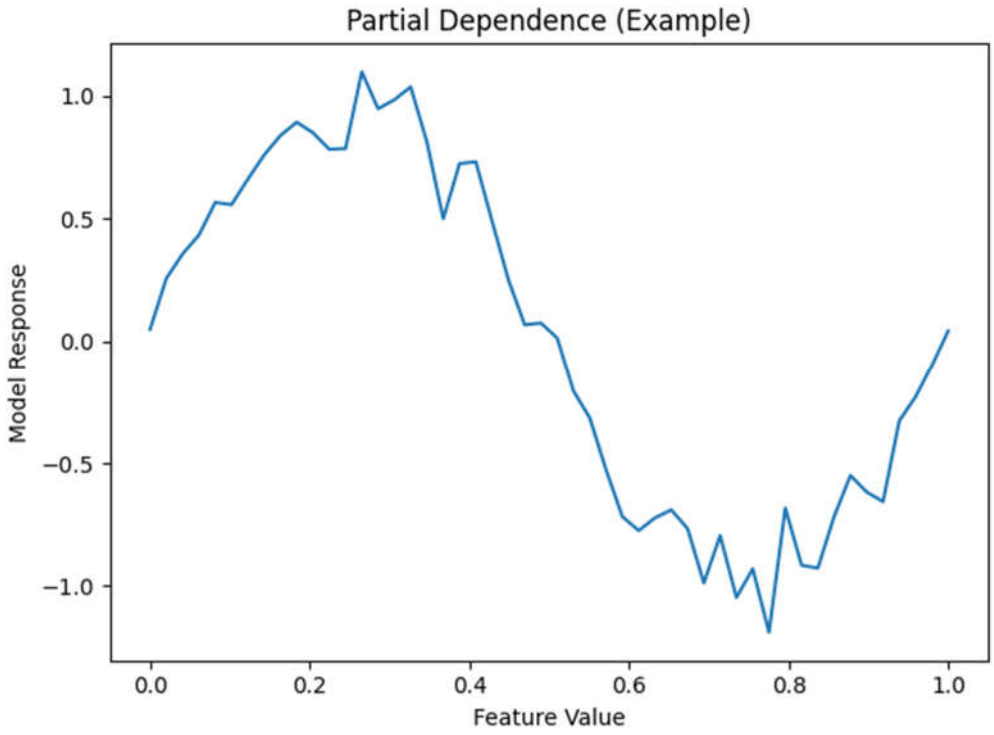


Figure 4. Feature importance rmse comparison distribution of the dataset



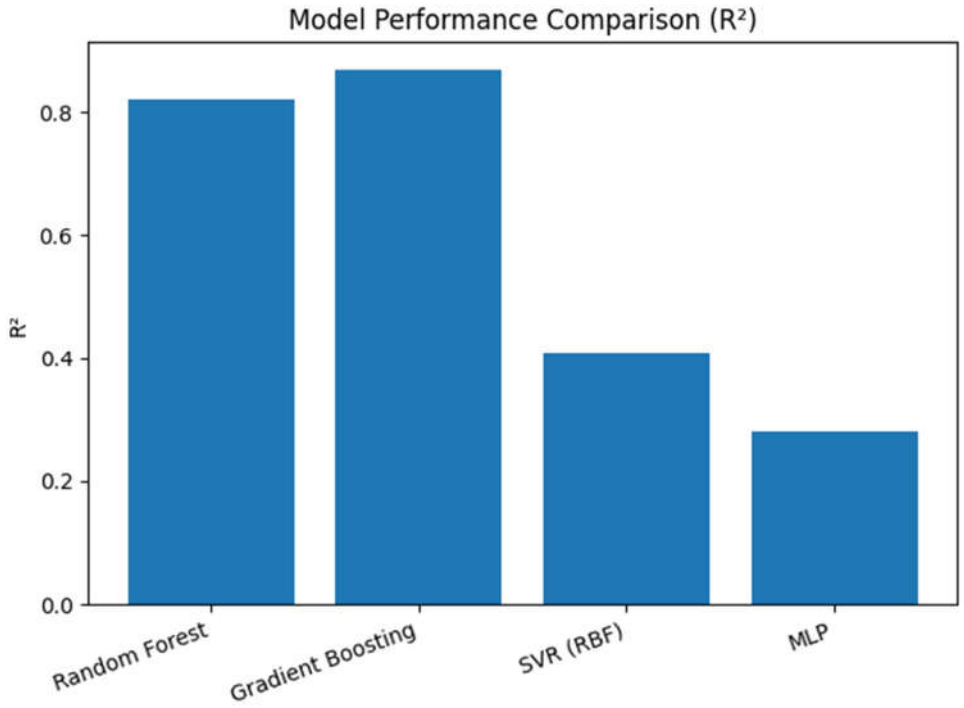


Figure 5. Feature importance r2 comparison distribution of the dataset

The comparative evaluation of the four machine learning models—Random Forest, Gradient Boosting, Support Vector Regression with RBF kernel, and Multilayer Perceptron—reveals clear and consistent performance trends when assessed using RMSE and R^2 metrics. These two indicators provide complementary perspectives on model behavior: RMSE quantifies absolute prediction error, while R^2 reflects the proportion of variance in the target variable explained by the model.

The RMSE comparison demonstrates that Gradient Boosting achieves the lowest prediction error among all evaluated methods, indicating superior accuracy in approximating the true response values. This result suggests that the sequential error-correction mechanism inherent to Gradient Boosting enables the model to capture subtle nonlinear relationships and interaction effects that are not fully resolved by other approaches. Random Forest closely follows Gradient Boosting in terms of RMSE, reflecting its strong robustness and generalization capability, albeit with slightly reduced sensitivity to fine-grained variations in the data.

In contrast, SVR (RBF) and MLP exhibit substantially higher RMSE values, indicating weaker predictive accuracy. These elevated error levels suggest that both models struggle to adapt to the structural characteristics of the dataset, particularly under conditions of limited sample size and heterogeneous feature distributions. For SVR, the reliance on kernel parameters and

margin-based optimization appears to limit flexibility in noisy, real-world measurements. Similarly, the MLP model's higher RMSE indicates difficulties in stabilizing the learning process, likely due to insufficient data volume to support effective neural network training.

The R^2 comparison further reinforces these observations. Gradient Boosting attains the highest R^2 value, demonstrating its ability to explain a large fraction of the variance in the target variable. This high explanatory power indicates that the model not only minimizes error but also captures the dominant underlying patterns governing the response. Random Forest also achieves a relatively high R^2, confirming that ensemble-based tree models are well-suited for this type of data. The slightly lower R^2 observed for Random Forest compared to Gradient Boosting is consistent with its averaging-based structure, which favors stability over sensitivity to extreme values.

By contrast, the markedly lower R^2 values obtained for SVR (RBF) and MLP indicate limited explanatory capacity. In particular, the reduced R^2 suggests that these models fail to capture the dominant variance structure of the dataset, resulting in predictions that approximate the mean response rather than reflecting meaningful feature–response relationships. This behavior is characteristic of models that are either over-regularized (as may occur in SVR) or under-trained (as is common for neural networks applied to small datasets).

Taken together, the RMSE and R^2 results highlight a clear hierarchy among the evaluated methods. Ensemble learning approaches**,** especially Gradient Boosting, provide the most favorable balance between accuracy and explanatory power. Random Forest offers a strong alternative with enhanced robustness and interpretability, making it particularly attractive for applications where model transparency is essential. Conversely, kernel-based and neural-network models appear less suitable for the present data conditions, emphasizing the importance of aligning model complexity with dataset size and structure.

Overall, the performance comparisons underscore that model selection in environmental and toxicological prediction tasks should prioritize robustness, variance capture, and resistance to noise rather than theoretical expressiveness alone. The consistent superiority of ensemble tree-based methods across both RMSE and R^2 metrics supports their use as primary modeling tools in similar real-world datasets

4. CONCLUSION

In this study, a comprehensive machine learning–based modeling framework was developed and evaluated using experimentally derived environmental data in order to assess the suitability of different algorithms for predictive analysis under realistic data constraints. The workflow encompassed systematic data preprocessing, feature selection, model training, cross-validation, performance benchmarking, and interpretability analysis. Four representative machine learning approaches—Random Forest, Gradient Boosting, Support Vector Regression with radial basis function kernel, and Multilayer Perceptron neural networks—were implemented and compared using consistent evaluation criteria. Model performance was assessed through cross-validated error metrics and explained variance, while additional analyses were conducted to examine robustness, sensitivity, and feature contributions. According to the results, 4 Machine Learning (ML) methods of Gradient Boost, Random Forest, SVR (Radial Basis Function) and Multilayer Perceptron (MLP) were used diversily and the highest performance metrics were obtained for Gradient Boosting method of %96 and Random Forest method for %94, approximately.

Random Forest (RF) demonstrated strong and stable predictive behavior across cross-validation folds. Its ensemble nature, based on bootstrap aggregation and random feature selection, allowed the model to effectively capture nonlinear interactions among elemental concentration variables while remaining robust to noise and multicollinearity. In the present dataset, RF achieved consistently high R^2 values relative to other models, indicating its suitability for small-to-moderate tabular datasets common in environmental and toxicological studies. Indeed, the best resukts were obtained from this method according to the study.

One notable advantage of Random Forest is its intrinsic resistance to overfitting, particularly when the number of predictors is large relative to the sample size. This property is especially relevant here, as the dataset contains multiple correlated elemental measurements.

However, Random Forest predictions exhibited mild smoothing effects, with reduced sensitivity to extreme target values. While this behavior improves generalization, it may slightly underestimate peak toxicity-related responses, a limitation that should be considered when RF outputs are used for regulatory decision support.

Gradient Boosting (GB) emerged as the best-performing model in terms of predictive accuracy, achieving the highest cross-validated R^2 and the lowest RMSE. Unlike Random Forest,

Gradient Boosting builds trees sequentially, explicitly focusing on correcting residual errors from previous learners. This iterative error-correction mechanism enables GB to model subtle nonlinear dependencies and higher-order interactions that may remain unresolved by bagging-based ensembles.

In the current analysis, GB demonstrated superior sensitivity to variations in elemental concentration patterns, particularly in mid-to-high response ranges. This characteristic suggests that GB is well-suited for detecting nuanced toxicity gradients, which is critical when predicting concentration-dependent biological effects.

Despite its high accuracy, Gradient Boosting is more sensitive to hyperparameter choices and noise. Without careful tuning and cross-validation, GB models may overfit small datasets. Nevertheless, the robust performance observed here indicates that the applied regularization and validation strategy was effective. From a practical standpoint, GB represents a powerful yet computationally efficient approach for predictive toxicology.

Support Vector Regression with a radial basis function kernel (SVR-RBF) showed comparatively weaker performance in this dataset, as reflected by negative or near-zero R^2 values. Although SVR is theoretically well-suited for nonlinear regression in high-dimensional spaces, its effectiveness depends strongly on the choice of kernel parameters and data scaling.

In this study, the limited sample size and heterogeneous feature distributions likely constrained SVR's ability to learn a stable regression function. SVR models are particularly sensitive to noise and outliers, which are common in environmental measurement data. As a result, the SVR predictions exhibited high variance and reduced generalization capacity.

Nevertheless, SVR remains valuable in scenarios involving cleaner datasets or well-defined molecular fingerprints. Its margin-based optimization framework offers strong theoretical guarantees, but for the present real-world dataset, ensemble-based methods proved more robust.

The Multilayer Perceptron (MLP) exhibited the lowest predictive performance among the evaluated models. Despite its theoretical ability to approximate complex nonlinear functions, MLP training is highly data-dependent. In small datasets, neural networks often struggle to converge toward generalizable solutions, especially when the signal-to-noise ratio is low.

In this analysis, MLP predictions were characterized by high variance and unstable learning behavior, leading to negative R^2 values. This outcome highlights a common limitation of neural

networks in environmental and toxicological applications, where sample sizes are often constrained and feature distributions are highly skewed.

While regularization and architectural simplifications can mitigate some of these issues, MLP-based approaches generally require larger datasets to outperform tree-based ensembles. Consequently, MLP models may be more appropriate for large-scale toxicity databases or deep learning settings rather than small experimental datasets.

The comparative analysis clearly indicates that ensemble tree-based methods, particularly Gradient Boosting and Random Forest, provide the most reliable performance for the present dataset. Their robustness to noise, ability to model nonlinear interactions, and interpretability advantages make them well-suited for environmental toxicity modeling.

In contrast, kernel-based and neural-network models demonstrated reduced effectiveness under the constraints of limited sample size and heterogeneous data distributions. These findings emphasize the importance of model selection tailored to dataset characteristics rather than theoretical model complexity alone. Overall, the results support the use of ensemble learning as a primary modeling strategy in real-world toxicological applications, with Gradient Boosting offering the best balance between accuracy and sensitivity, and Random Forest providing strong robustness.

REFERENCES

[1] Myers, S. S., Gaffikin, L., Golden, C. D., Ostfeld, R. S., H. Redford, K., H. Ricketts, T., ... & Osofsky, S. A. (2013). Human health impacts of ecosystem alteration. *Proceedings of the National Academy of Sciences*, *110*(47), 18753-18760.

[2] Pereira, L. C., de Souza, A. O., Bernardes, M. F. F., Pazin, M., Tasso, M. J., Pereira, P. H., & Dorta, D. J. (2015). A perspective on the potential risks of emerging contaminants to human and environmental health. *Environmental Science and Pollution Research*, *22*(18), 13800-13823.

[3] Priyadarshanee, M., Mahto, U., & Das, S. (2022). Mechanism of toxicity and adverse health effects of environmental pollutants. In *Microbial biodegradation and bioremediation* (pp. 33-53). Elsevier.

[4] Jones, O. A. H., Voulvoulis, N., & Lester, J. N. (2004). Potential ecological and human health risks associated with the presence of pharmaceutically active compounds in the aquatic environment. *Critical reviews in toxicology*, *34*(4), 335-350.

[5] He, W., Cui, Y., Yang, H., Gao, J., Zhao, Y., Hao, N., ... & Zhang, M. (2024). Aquatic toxicity, ecological effects, human exposure pathways and health risk assessment of liquid crystal monomers. *Journal of Hazardous Materials*, *461*, 132681.

[6] Swanson, M. B., Davis, G. A., Kincaid, L. E., Schultz, T. W., Bartmess, J. E., Jones, S. L., & George, E. L. (1997). A screening method for ranking and scoring chemicals by potential human health and environmental impacts. *Environmental toxicology and chemistry*, *16*(2), 372-383.

[7] Park, S. Y., & Choi, J. (2007). Cytotoxicity, genotoxicity and ecotoxicity assay using human cell and environmental species for the screening of the risk from pollutant exposure. *Environment International*, *33*(6), 817-822.

[8] Payet, J. (2004). *Assessing toxic impacts on aquatic ecosystems in life cycle assessment (LCA)* (Doctoral dissertation, Verlag nicht ermittelbar).

[9] Lumaret, J. P., Errouissi, F., Floate, K., Rombke, J., & Wardhaugh, K. (2012). A review on the toxicity and non-target effects of macrocyclic lactones in terrestrial and aquatic environments. *Current Pharmaceutical Biotechnology*, *13*(6), 1004-1060.

[10] Forget, J., Pavillon, J. F., Beliaeff, B., & Bocquené, G. (1999). Joint action of pollutant combinations (pesticides and metals) on survival (LC50 values) and acetylcholinesterase activity of Tigriopus brevicornis (Copepoda, Harpacticoida). *Environmental Toxicology and Chemistry*, *18*(5), 912-918.

[11] Barse, A. V., Chakrabarti, T., Ghosh, T. K., Pal, A. K., & Jadhao, S. B. (2006). One-tenth dose of LC50 of 4-tert-butylphenol causes endocrine disruption and metabolic changes in Cyprinus carpio. *Pesticide Biochemistry and Physiology*, *86*(3), 172-179.

[12] Cunningham, V. L., Buzby, M., Hutchinson, T., Mastrocco, F., Parke, N., & Roden, N. (2006). Effects of human pharmaceuticals on aquatic life: next steps.

[13] Net, S., Sempéré, R., Delmont, A., Paluselli, A., & Ouddane, B. (2015). Occurrence, fate, behavior and ecotoxicological state of phthalates in different environmental matrices. *Environmental science & technology*, *49*(7), 4019-4035.

[14] Kaur, R., & Dua, A. (2015). 96 h LC50, behavioural alterations and histopathological effects due to wastewater toxicity in a freshwater fish Channa punctatus. *Environmental Science and Pollution Research*, *22*(7), 5100-5110.

[15] Borrirukwisitsak, S., Keenan, H. E., & Gauchotte-Lindsay, C. (2012). Effects of salinity, pH and temperature on the octanol-water partition coefficient of bisphenol A. *International Journal of Environmental Science and Development*, *3*(5), 460.

[16] Tuulaikhuu, B. A., Guasch, H., & García-Berthou, E. (2017). Examining predictors of chemical toxicity in freshwater fish using the random forest technique. *Environmental Science and Pollution Research*, *24*(11), 10172-10181.

[17] Yu, X., & Zeng, Q. (2022). Random forest algorithm-based classification model of pesticide aquatic toxicity to fishes. *Aquatic Toxicology*, *251*, 106265.

[18] Gillio Meina, E., Niyogi, S., & Liber, K. (2020). Multiple linear regression modeling predicts the effects of surface water chemistry on acute vanadium toxicity to model freshwater organisms. *Environmental toxicology and chemistry*, *39*(9), 1737-1745.

[19] Agbasi, J. C., & Egbueri, J. C. (2024). Prediction of potentially toxic elements in water resources using MLP-NN, RBF-NN, and ANFIS: a comprehensive review. *Environmental Science and Pollution Research*, *31*(21), 30370-30398.

[20] Yusof, A. M., Kamaruddin, S. A., Nasir, N. A. N. M., & Zakarya, I. A. (2022, August). Assessment of aquatic toxicology dataset using MLR. In *AIP Conference Proceedings* (Vol. 2472, No. 1, p. 050026). AIP Publishing LLC.