Developing machine learning algorithms to predict the dissolution of zinc oxide nanoparticles in aqueous environment

Ntsikelelo Yalezo [a], Ndeke Musee [b], Michael O. Daramola [a,*]

[a] Department of Chemical Engineering, University of Pretoria, Private Bag X20, Hatfield 0028, Pretoria, South Africa
[b] Former employee in Department of Chemical Engineering at University of Pretoria, South Africa

Keywords:
Machine learning
nZnO dissolution
Surface transformation
Aqueous environment
Meta-analysis

A B S T R A C T

Engineered nanoparticles (ENPs) are of particular concern due to their ubiquitous occurrence and potential to cause adverse effects on aquatic biota. Consequently, a comprehensive understanding of ENP interactions and the mechanisms that underpin their fate and behaviour in the aquatic system is important to support their long-term applications and protection of ecology. However, due to a wide range of physicochemical parameters, as well as possible dynamic interactions with natural colloid particles, it is not practical to undertake experimental testing for each variation of ENPs using different aquatic permutations. This study describes machine learning (ML) algorithms for prediction of nZnO dissolution in aquatic systems using experimental data. The input parameters with the highest correlation were size and pH. On the contrary, categorical input variables such as coating, coating type, salt, and NOM type had a low correlation. The random forest regression and the extreme gradient boost algorithms performed remarkably well, with coefficients of determination (R^2) of 0.85 and 0.92, respectively. The least effective method was multiple linear regression, which had a root mean square error of 0.15 and an R^2 of 0.31. ML offers a convenient and low-cost approach for screening nZnO dissolution in aquatic systems.

1. Introduction

Zinc oxide nanoparticles (nZnO) have received significant attention in various areas ranging from automotive, biomedical, energy, and electronic products (Foss Hansen et al., 2016; Grillo et al., 2018; Sengul and Asmatulu, 2020). This is due to their distinct features including a wide band gap of 3.37 electron volts (Debanath and Karmakar, 2013), antimicrobial activity (Sirelkhatim et al., 2015), piezoelectric and pyroelectric properties (Parihar et al., 2018). Apart from the progressive application across various product categories, the widespread use of nZnO results in ubiquitous occurrence in aquatic environments and potentially deleterious effects on aquatic biota, i.e. bacteria, plants, microorganisms, etc., (Leareng et al., 2020: Schiavo et al., 2016). For these reasons, a thorough elucidation of the behaviour and interactions of ENPs with natural colloids in aquatic environments is necessary to address the existing environmental safety concerns.

ENPs in aquatic systems are subjected to various transformation processes. These processes have a profound influence on their adverse effects (Abbas et al., 2020). Dissolution is one of the important and highly studied chemical processes that impact nZnO biodurability and bioaccumulation (Hou et al., 2018; Mahaye et al., 2017; Musee et al., 2014). For example, a high dissolution of nZnO into Zn^{2+} is concomitant with enhanced deleterious effects, as demonstrated by studies conducted against bacteria and microorganisms such as *Bacillus subtilis* (Leareng et al., 2020) and *Escherichia coli* (Li et al., 2011), respectively. In contrast, reduced concentration of Zn^{2+} as the result of rapid aggregation of nZnO in aqueous media is concomitant with high attachment efficiency and reduced toxicity effects, as demonstrated by studies conducted on *Bacillus subtilis* (Leareng, et al., 2020), *Lemna minor* plant (Chen et al., 2016), and microalgae *Dunaliella tertiolecta* (Schiavo et al., 2016).

Traditional to quantify the dissolution, dissolution rate, and reaction kinetics of solid surfaces, including ENPs in aqueous media, mathematical equations such as Noyes-Whitney, Nernst-Brunner, and Hixson-Crowell (Siepmann and Siepmann, 2013) and numerically derived zero-, first-, or second-order reaction equations (Utembe et al., 2015) have been reported. Despite the advantages of these modelling concepts, they have several drawbacks such as being time-consuming, requiring complex calculations, only taking into account one or a few parameters at a time, and frequently being relevant to spherical nanoparticles (Song et al., 2023). As a result, the dynamic interactions of ENPs with inorganic ions and natural colloids, such as NOM are not adequately

* Corresponding author.
E-mail address: michael.daramola@up.ac.za (M.O. Daramola).

reflected; despite these interactions have a significant impact on toxicity.

Additionally, to date, a large amount of data has been generated in various experimental research studies to comprehend the influence of physicochemical (e.g., size, shape) and water chemistry properties toward the dissolution of ENPs (see Table S2). However, it is challenging to interpret and deduce significant trends from the extremely varied data. This is partly because of the vast variety of physicochemical parameters that affect ENP behaviour, as well as the considerable variability of the exposure media, due to non-standardised experimental protocols (Ban et al., 2018; Basei et al., 2019). As a result, the use of data-driven methods such as machine learning (ML) is, therefore, required as the volume of data increases to provide a clear understanding of the interactions and mechanisms underlying the dissolution of ENPs.

ML is a branch of artificial intelligence that learns from data without explicitly being programmed (Findlay et al., 2018; Papa et al., 2015; Sizochenko et al., 2019). The use of ML has attracted growing interest in several scientific domains, including nanotoxicology and nanoecotoxicology, in recent years (Ban et al., 2020; Furxhi et al., 2019; Mirzaei et al., 2021; Peng et al., 2020; Takahashi and Takahashi, 2019). ML algorithms have demonstrated an enormous ability to predict the adverse effects of ENPs on numerous aquatic organisms, including *Escherichia coli (E. coli)* (Fjodorova et al., 2017), and *Daphnia magna* (Balraadjsing et al., 2022) based on experimentally measured or computed physicochemical parameters.

Apart from the remarkable predictive capabilities demonstrated by ML models, additionally, the approach has been utilised for a variety of interrelated purposes including identifying scientific knowledge and underlying patterns (Balraadjsing et al., 2022; Concu et al., 2017; Hou et al., 2020). Recently, results of the ML algorithms showed input parameters of zeta potential (ζ), pH, and time as good predictors for estimating the dynamic aggregation of nZnO and titanium dioxide ($nTiO_2$) in freshwater systems (Yalezo and Musee, 2023). Elsewhere, ML analysis of relative attribute importance showed attributes of exposure interval, aggregation size, dosage, and formation enthalpy, as predominant variables for predicting the toxicity of various ENPs (nZnO, $nTiO_2$, silicon dioxide ($nSiO_2$), aluminium oxide (nAl_2O_3), copper oxide (nCuO) and iron oxide (nFe_2O_3) (Choi et al., 2018). As a result, ML can guide future experimental investigations by narrowing the focus from many predictors that are concomitant with complexity to a smaller number of variables.

However, so far despite the dissolution of ENPs being recognised as an essential process that influences bioavailability and bioaccumulation; ML methods, on the other hand, are lacking for data mining or predicting ENP dissolution in aquatic systems. This work, describes a range of ML algorithms of artificial neuron network (ANN), support vector regression (SVM), multiple linear regression (MLR), random forest regression (RFR), and extreme gradient boosting (XGBoost) to elucidate properties that can be used as precursors for evaluation and screening of nZnO dissolution in the aquatic environment.

2. Materials and methods

2.1. Overview of the modelling process

The field of nanoecotoxicology has seen a steady increase in experimental data recently; yet, for two reasons, the data is considered information- and knowledge-poor. First, there are discrepancies in the reporting protocols in the accessible studies concerning exposure media, high instrument variability to measure different parameters and inadequate controls. The second issue is the paucity of uniform and structured datasets for efficient analysis (Basei et al., 2019; Trinh et al., 2018). To address the lack of readily available quantitative datasets, secondary data published from a variety of experimental sources was extracted following the *meta*-analysis procedure depicted in Fig. 1.

Meta-analysis is a quantitative analysis procedure that uses secondary data obtained from literature reviews (Foley et al., 2018). The benefits of *meta*-analysis include reducing noise and bias observed in individual research and combining pertinent studies to capture all conceivable permutations that drive the issue of interest (Gurevitch et al., 2018). The *meta*-analysis was conducted following the "Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA)" guidelines (Deji et al., 2021). Section 2.2 provides a summary of criteria applied for the data collection process and identification of input–output vectors; Section 2.3 discusses the pre-processing of the continuous and categorical parameters; Sections 2.4 and 2.5 provide details on the training of a range of ML algorithms and evaluation criteria, respectively.

2.2. Data extraction

A comprehensive search of the literature in online databases, including PubMed, Google Scholar, ScienceDirect, American Chemical Society, SpringerLink, and Web of Science, yielded research studies on the dissolution of nZnO. Boolean logic operators (AND, OR, and NOT) were used to search for keywords that appeared alone or in combination, such as (dissociation OR ions release) AND (dissolution OR surface transformation) AND (nanoparticles OR nanomaterial), AND (aqueous media OR natural water OR freshwater), among others. Fig. 2 displays the density visualisation map of key terms generated using the VOSviewer program (https://www.vosviewer.com/).

A total of 52 researched studies matched the keywords and the selected studies depicted in Table S2 were carefully chosen based on the

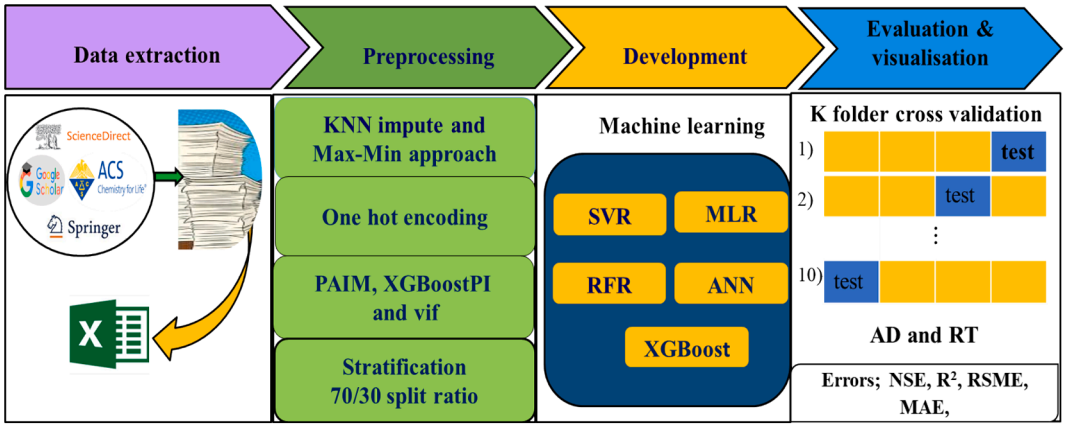

Fig. 1. Workflow depicting the modelling process: data extraction, preprocessing, model development, evaluation, and visualisation.

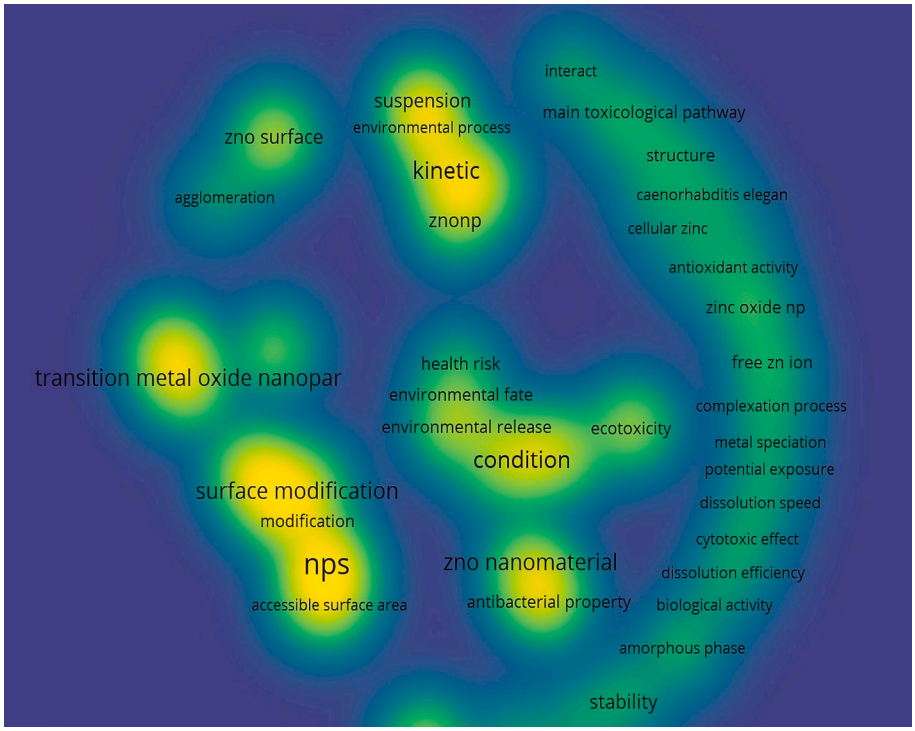

Fig. 2. Density visualisation map showing the co-occurrence of the keywords. The blue colour denotes low density, and the contrary holds for yellow.

following set of criteria, namely; (i) nZnO as the ENP of interest and (ii) freshwater (i.e., rivers, lakes, etc.) and/or aqueous systems that detail the effects of water chemistry and physicochemical properties. The bibliographies of the retrieved articles were searched for references that were not found in the electronic search. Furthermore, using accessible online software like GetData Graph Digitiser or Plot Digitiser (https://getdata-graph-digitizer.com/), the dissolution data of nZnO in the aquatic environment were extracted from peer-reviewed publications (Gagliardi et al., 2016; Wang et al., 2017; Yalezo and Musee, 2023). The inputs included the concentration of nZnO concentration (X_1, mg/ℓ), duration (X_2, h), NOM (X_3, mg/ℓ), shape (X_4), IS (X_5, mM), size (X_6, nm), pH (X_7, dimensionless), NOM type (X_8), coating (X_9), coating type (X_{10}), salt type (X_{11}) and zeta potential (ζ) (X_{12}, mv). The concentration of Zn^{2+} (mg/ℓ) was assigned as the model output.

2.3. Pre-processing

2.3.1. Missing data and normalisation

Missing data poses a significant challenge in the application of ML (Balraadjsing et al., 2022; Mirzaei et al., 2021; Sizochenko et al., 2019). Missing data can be categorised as missing completely at random (MCAR): missing data that are not interrelated to any other parameter, missing at random (MIR): missing data that are associated with other parameters, and not missing at random (NMAR): the missing data that are connected to the parameter itself (Batista and Monard, 2003; Troyanskaya et al., 2001). In this work, the KNN impute was applied to populate missing data (Yalezo and Musee, 2023). The distance between neighbours was calculated using the Euclidean distance (p = 2). The number of neighbours investigated was in the range of 1 to 20. Theoretically, the smaller value of k results in overfitting, and the contrast holds for higher values (Zhang et al., 2017). Python 3.0 was used to implement KNN imputation in the Scikit-Learn Library. In addition, the input data values were scaled between 0 and 1 to reduce the variability of the raw original data using Equation (1):

$$y_j = \frac{x_j - \min x_i}{\max x_i - \min x_i} \quad (1)$$

where, j= 1,2…m; i=1,2,…,n, x_i is the original data sequence, y_j is the pre-processed data, $\min x_i$ and $\max x_i$ denote the smallest and the largest values of the sequence x_i, respectively, and x_j is the data point normalised.

2.3.2. One hot encoding

The ML-based predictive algorithms are generally developed from numerical data pairs $(x_1, y_1, \ldots, x_j, y_j,)$ where input vectors $x = (x_1, x_2, \ldots, X_n)^K$, y_j is the corresponding output, j = 1, 2,…, *m*, and n =1, 2,…, k. Categorical features cannot be utilised by ML algorithms and thus must be transformed into numerical data before modelling (Findlay et al., 2018; Furxhi et al., 2019). The categorical variables in Table S1 were converted by applying a one-hot encoding approach. A one-hot encoding approach helps to create multiple binary numerical features or dummy variables 'feature vectors' from categorical features (Balraadjsing et al., 2022; Glaubitz et al., 2022).

2.3.3. Feature analysis

Feature analysis has multiple objectives, including (i) determining collinearity or multicollinearity and (ii) ranking variables and/or reducing dimensions (Bahl et al., 2019; Subramanian and Palaniappan, 2021). Multicollinearity or collinearity occurs when two or more descriptors have a high correlation feature space, and this can result in data snooping and confounding. To investigate collinearity or multicollinearity, the variance inflation factor (vif) was computed using Equation (2) (Subramanian and Palaniappan, 2021). The vif score of 1.0 and < 5.0 indicated no correlation and moderate multicollinearity that pose a serious problem, respectively. The vif scores between 5 and 10 and scores > 10 indicated high, and severe multicollinearity, respectively.

$$vif_j = \frac{1}{1 - R_j^2} \quad (2)$$

Where R_j^2 represent the coefficient of determination of x_j ($j = 1, 2\ldots n$)

Furthermore, dimension reduction is the process of determining a subset or a minimal set of variables that can help construct a parsimony

model. There are several methods to reduce the curse of dimensionality and/or ranking variables. These include grey relation coefficients (GRC) (Wang et al., 2013), partial rank correlation coefficient (PRCC) (Khoshroo et al., 2018), MLR (Chen et al., 2019; Lee et al., 2016), random forest feature importance (RFFI) (Ban et al., 2018; Kerckhoffs et al., 2019), and principal component analysis (PCA). Spearman, Pearson, and PRCC have limited application confined to variables that have underlying linear relationships (Alimissis et al., 2018) – a phenomenon that is uncommon for variables that influence the transformations of ENPs in aqueous media. Permutation accuracy importance measurement (PAIM) previously discussed by Yalezo and Musee (2023) and XGBoost feature importance (XGBoostFI) were applied to investigate the importance and dimension reduction of the variables. The XGBoostFI and PAIM were implemented in Python v3.0 and the R software package, respectively.

2.3.4. Data splitting

The accuracy of the generated ML algorithm is highly dependent on the amount of the training data set (Subramanian and Palaniappan, 2021). There are two important aspects for choosing an appropriate split ratio: parameter estimations lead to increased variance with fewer training data. On the contrary, the performance statistic produces higher variance with fewer test data. It is important to partition the data in a way that prevents neither of these situations from occurring (Balraadjsing et al., 2022; Furxhi et al., 2019).

The most popular method is the Pareto principle, which employs an 80:20 ratio (Takahashi and Takahashi, 2019). More ratios have been reported including 90:10 (Findlay et al., 2018), 60:40 (Balraadjsing et al., 2022), and 70:30 (Papa et al., 2015). The ideal split ratio depends on the volume of data. For example, the split ratios of 70:30 and 60:40 are typically appropriate for sufficient representation within "smaller" datasets (n < 1000) (Yalezo and Musee, 2023). In this study, both random sampling and stratified splitting were applied using a split ratio of 70:30. The stratification approach promotes higher prediction quality, including balancing the distribution of multiple classes or groups (Glaubitz et al., 2022).

2.4. ML development

Five distinct ML algorithms – MLR, SVR, XGBoost, ANN, and RFR – were supervised to predict the dissolution of nZnO in an aqueous medium. Python 3.0 and the Scikit-Learn package were used to implement the ML algorithms. Table 1 summarises the hyper-parameters applied in the developed ML algorithms. Details on ANN, SVR, and RFR are covered in our previous study (Yalezo and Musee, 2023); for convenience, only XGBoost is discussed.

Table 1
Various hyper-parameters for different ML.

ML	Hyper-parameters
ANN	Loss function: [stochastic gradient descent (SGD), adaptive momentum (Adam)]
	Activation functions: [rectified linear unit (ReLU), sigmoid, tanh]
	Epochs:1000, neurons: 10, momentum: 0.09, and learning rate: 0.1
RFR	Criterion: [squared error]
	Maximum features: [sqrt, log2]
	Number of estimators: [20, 100, 200, 500]
	randomized state: [42]
SVR	Kernel: [radial basis function (RBF), polynomial (Poly)]
	C (cost of violation of the constraint): [1]
	ε (epsilon): [0.1, 0.3]
	γ (gamma): [1,10]
KNN	Number of neighbours: [1—20]
	Algorithm: [auto]
	Euclidean distance (p): [2]
XGB	learning rate: [0.001, 0.01, 0.05, 0.1, 0.2, 0.3]
	Number of estimators: [50, 100, 500, 1000]

XGBoost is widely applied for both regression and classification. An arbitrarily differential loss function and the gradient descent optimisation procedure are used to fit XGBoost models (Chen et al., 2015). XGBoost and gradient boosting (GB) use the same principle, but the former has regularisation to reduce overfitting and bias (Ogunleye and Wang, 2019). XGBoost is simple to use and understand, and it offers flexibility to manage small amounts of data. Furthermore, non-linearity, extreme findings, and feature prioritization are all possible with boosting models (Dong et al., 2022; Glaubitz et al., 2022). Suppose, **m** and **n** represent the number of samples and features, respectively, then the data set is defined; $\{(x_j,\ y_j)$: J =1, 2…*m*-dimensional space, $\in R^{n\ x\ m}$ and $y \in R^{m\ x\ 1})\}$. The mathematical function of XGBoost is expressed as follows:

$$y_j^{(dt)} = \sum_{k=1}^{K} f_k(x_j) \tag{3}$$

Where f_k is the independent tree, $F = (f_1,\ f_2, ..,\ f_k)$ denotes the regression trees, $y_j^{(dt)}$ is the estimated value of sample *j* after *K-th* iterations. The *dt* is the decision trees

The cost or loss function is expressed as follows:

$$L^{(dt)} = \sum_{j=1}^{m} l(y_j,\ y_j^{(dt)}) + \Omega(f_k)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\ \lambda \|w\|^2 \tag{4}$$

Where *l* is the deviation between the actual and forecasted values, *T* and *w* are the number and weights of the leaf nodes, respectively. In addition, the constants γ and λ are terms for regularisation. $\Omega(f_k)$ indicates the complexity of the model, and can be tuned to reduce overfitting, bias, or variance.

Subsequently, to minimise the loss function the expression is given by the following;

$$L^{(dt)} = \sum_{j=1}^{m} l[(y_j,\ y_j^{(dt-1)}) + f_{dt}(x_j)] + \Omega(f_{dt}) \tag{5}$$

In addition, the Taylor approximation of the loss function is expressed as follows:

$$L^{(dt)} = \sum_{j=1}^{m} l[g_j f_{dt}(x_j) + \frac{1}{2} h_j {f_{dt}}^2(x_j)] + \Omega(f_{dt}) \tag{6}$$

Where $h_j = \partial^2_{y_j^{(dt-1)}}(y_j,\ y_j^{(dt-1)})$ and $g_j = \partial_{y_j^{(dt-1)}} l\left(y_j,\ y_j^{(dt-1)}\right)$ are the second and first derivatives respectively.

The scoring function in Equation (7) evaluates the optimal weight value to compute the predicted value for each leaf node.

$$L^{(f)} = -\frac{1}{2}\sum_{n=1}^{T} \frac{{G_j}^2}{H_j + \lambda} + \gamma T \tag{7}$$

Where the I_i instance set of leaf *i*-th, $G_j = Æ©_{j\in I_i}\ g_j$ and $H_j = Æ©_{j\in I_i}\ h_j$

2.5. Model evaluation

The best-performing ML approach (es) was selected using k-fold cross-validation (CV). This process partitioned the calibration and validation data sets into several folders. The K-folder CV approach reduces over-representation and promotes a low-bias prediction under sparse and inadequate data. Many research studies make use of k = 5 or 10 (Findlay et al., 2018; Subramanian and Palaniappan, 2021). Low bias and variance are produced by high values of k (Balraadjsing et al., 2022; Ban et al., 2020; Choi et al., 2018). The data was divided into ten equivalent subsamples, nine for training purposes, and one for validation. In addition, several statistical measures were used to evaluate ML models. These include the coefficient of determination (R^2), Nash-Sutcliffe efficiency (NSE), root mean squared error (RMSE), and mean

absolute error (MAE), as described in Equations (8) to (11), respectively. The values of R^2 and NSE close to one, and values of RMSE and MAE close to zero signify better model performance (Alexander et al., 2015).

$$R^2 = \left(\frac{\sum_i^k (t_i - F)(y_i - T)}{\sqrt{\sum_i^k (t_i - F)^2} \bullet \sqrt{\sum_i^k (y_i - T)^2}} \right)^2 \quad (8)$$

$$NSE = \left(1 - \frac{\sum_{i=1}^k (t_i - y_i)^2}{\sum_{i=1}^k (t_i - F)^2} \right) \quad (9)$$

$$RMSE = \sqrt{\frac{\sum_i^k (t_i - y_i)^2}{k}} \quad (10)$$

$$MAE = \frac{\sum_i^k |(t_i - y_i|}{k} \quad (11)$$

Where y_i is the projected output, k is the quantity of samples, T is the average of the projected output, t_i and F are the actual output, and the mean of the actual output, respectively (Yalezo and Musee, 2023).

2.5.1. Randomisation test

The randomisation test (RT) is a procedure that permits the determination of whether the observed data are different from the random distribution generated by shuffling the observed data (Ojala and Garriga, 2010). To pass the RT, we consider a null (H_0) and alternative hypothesis (H_A) as described in Equations 12 and 13, respectively. A variety of test statistics, such as mean and variance, can be used; in this instance, the R^2 was selected (Ciszewski et al., 2024). To reject H_0 or otherwise, the probability significance level was based on alpha (α) of 5% as the confidence level. If $p < \alpha$, we reject H_0 and the developed models are randomly generated. Otherwise, If $p \geq \alpha$, then we do not have sufficient evidence to reject the H_0. Meaning, that the random sample belongs to the distribution of permutated results (Valente et al., 2021).

H_0 : distribution of permutated samples = random sample (12).

H_A : distribution of permuted samples ≠ random sample (13).

2.5.2. Applicability domain

The applicability domain (AD) is a concept that provides essential information regarding the endpoint that is predicted, the model algorithm used, the scope of the model and associated limitations, model performance and properties of the model descriptors of the training set (Hanser et al., 2016). AD can be characterised using a variety of techniques, including distance-based methods, probability density distribution methods, ranges of response variables, and geometric methods, among others (Li et al., 2022). Since the probability density distribution approach is regarded as one of the best AD measures to produce good performance, it was employed in this study. The density at point x is defined as the fraction d (x/h) of the data values per unit of measurement that fall in an interval h.

3. Results and discussion

3.1. Data analysis

A total of 791 data points were extracted on nZnO dissolution from publications that are summarised in Table S2. The dataset initially had continuous (n=7) and categorical (n=5) features namely; nZnO concentration (X_1, mg/ℓ), duration (X_2, h), NOM (X_3, mg/ℓ), shape (X_4), IS (X_5, mM), size (X_6, nm), pH (X_7, dimensionless), NOM type (X_8), coating (X_9), coating type (X_{10}), salt type (X_{11}) and zeta potential (ζ) (X_{12}, mv). The concentration of Zn^{2+} (X_{13}, mg/ℓ) was used as a model output. The standard deviations (SD), means, percentages of missing values of the input variables of IS, shape, ζ, NOM, nZnO concentration, and time are indicated in Table 2. In Fig. S1, the KNN model with k = 9 showed the best performance, as such it was used for the imputation of the missing values. The shape and ζ had insufficient data and these inputs were removed, as they were likely to result in data snooping.

3.2. Feature analysis

Variable selection helps avoid the dimensionality curse, reduces bias or noise, and improves model generalisation. The results described in Fig. 3 did not show existing multi-collinearity as the vif scores were less than 5. In addition, PAIM and XGBoostFI results in Fig. 4 for the input parameters of time, NOM, nZnO concentration, size, IS, and pH demonstrated a correlation of greater than 0.25 with the Zn^{2+} concentration. Consequently, these variables were identified as significant. The input variable of size had the highest correlation with the concentration of Zn^{2+}. Categorically input variables such as coating, salt type, coating type, and NOM type showed PAIM and XGBoostFI coefficients closer to zero (< 0.07). These variables were less significant for the prediction of the Zn^{2+} concentration. These findings are consistent with previously reported ML results by Goldberg et al. (2015). For example, a research study by Goldberg et al. (2015) showed that the qualitative variables, including the type of NOM, salt, and coating, had low significance or importance in determining the influence parameters of ENP transport–retained fraction and retention profiles – in saturated columns.

However, while these findings show consistency with the modelling results reported by Goldberg et al. (2015), they appear to contradict the experimental suggestions that the NOM type, electrolytes, and coating types also play a significant impact on transformation processes (Abbas et al., 2020; Louie et al., 2016, 2013). For example, in experimental literature studies, it has been demonstrated that the types of electrolytes impact the transformation processes of ENPs differently in aquatic systems (Chowdhury et al., 2012). In addition, coating and coating agents including citrate and polyvinylpyrrolidone have been observed to reduce the dissolution rate because of the shielding effect compared to bare ENPs (Lodeiro et al., 2016; Sharma et al., 2014).

To account for low PAIM and XGBoostFI coefficients in Fig. 4 of categorical variables, the concept of causal *versus* correlation, as well as the prediction *versus* significance concept were considered (Lo et al., 2015). Causation indicates a variation in one or more variables that results in the same effect on other variables. In contrast, correlation is a statistical metric that describes the relationship between two or many variables. A change in one variable does not automatically cause the same effect on the other variables (Ni et al., 2017; Shipley, 2016). As covered in our previous study (Yalezo and Musee, 2023), the experimentally reported significant variables are not always good predictor variables. Prediction is influenced by correlation as opposed to causal

Table 2
Type of data with missing percentages, and descriptive statistics for input variables.

Variables	Units	Type of inputs	Missing data (%)	Mean	SD
NOM	$mg \bullet l^{-1}$	Continuous	5.03	0.03	0.09
pH	–	Continuous	–	0.62	0.14
IS	mM	Continuous	1.21	0.17	0.24
Size	nm	Continuous	–	0.13	0.34
nZnO concentration	$mg \bullet l^{-1}$	Continuous	1.07	0.19	0.26
Time	Hour	Continuous	2.03	0.11	0.24
Shape	–	Categorical	50.3	–	–
ζ	mV	Continuous	62.4	0.12	0.17
NOM type	–	Categorical	–	–	–
Salt type		Categorical	–	–	–
Coating	–	Categorical	–	–	–
Coating type	–	Categorical	–	–	–
Zn^{2+} concentration	$mg \bullet l^{-1}$	Continuous	–	0.24	0.21

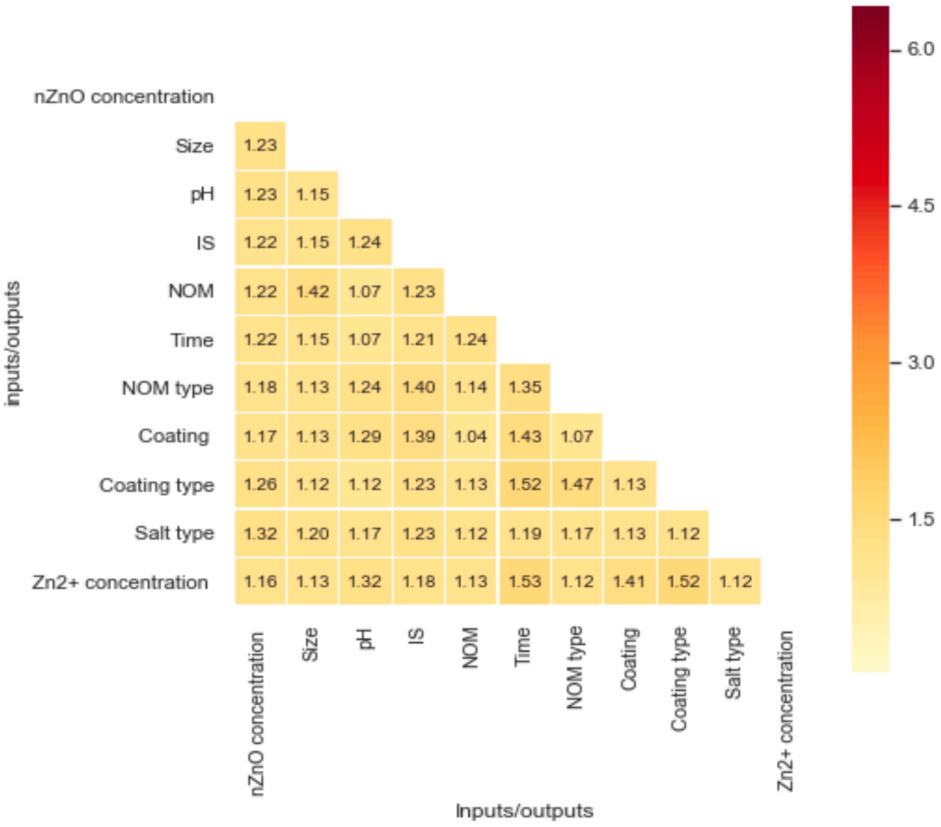

Fig. 3. Vif values to estimate the multi-collinearity.

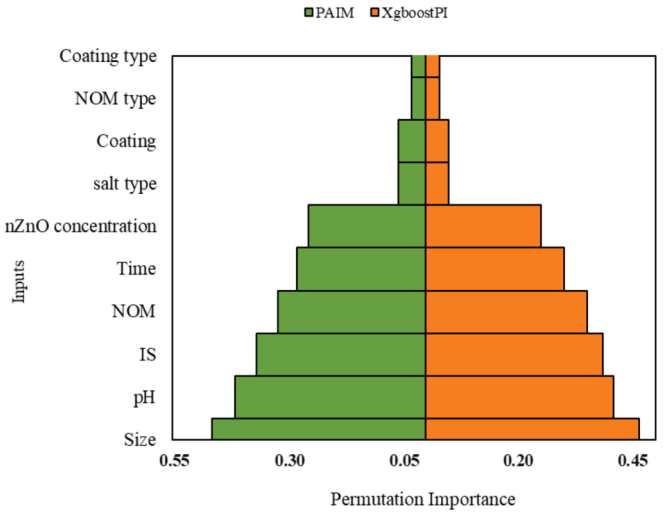

Fig. 4. Bipyramid diagram depicting feature significance results using both PAIM and XGBoostFI.

effect; hence, experimentally reported significant variables do not automatically possess high predictive power.

3.3. Selecting optimisers for different ML algorithms

Incorporating only the variables with high correlation provides a simpler, parsimonious predictive model with enhanced performance and generalisation capability (Subramanian and Palaniappan, 2021). To investigate the best optimisers for different ML algorithms the input variables of size, pH, IS, NOM, time, and nZnO concentration were utilised.

3.3.1. Artificial neural network

Results in Table 3 were developed based on the ANN architecture with 11 neurons, which had the least MAE in Fig. 5a. The best model was ANN3 with the Adam and ReLU functions and metric values of R = 0.82 and RMSE = 0.13. These outcomes differ from earlier research by Yalezo

Table 3
Performance parameters of the prediction models on the dissolution of nZnO for the training and testing sets.

				RMSE		R	
Model		Combination		Train	Test	Train	Test
ANN	1	Adam	Tanh	0.12	0.19	0.69	0.62
	2		Sigmoid	0.13	0.18	0.60	0.55
	3		ReLU	0.10	0.13	0.88	0.82
	4	SGD	Tanh	0.19	0.21	0.62	0.56
	5		Sigmoid	0.20	0.22	0.55	0.49
	6		ReLU	0.19	0.21	0.63	0.59
RFR	1	Trees	$20^{a,b}$	0.09	0.12	0.80	0.71
	2		$100^{a,b}$	0.03	0.06	0.97	0.92
	3		$200^{a,b}$	0.15	0.19	0.89	0.73
	4		$500^{a,b}$	0.17	0.20	0.89	0.71
SVR	1	Rbf	$(1^{c}, 0.1^{d},1^{e})$	0.10	0.12	0.79	0.70
	2		$(1^{c}, 0.3^{d},1^{e})$	0.09	0.10	0.90	0.87
	3		$(1^{c},0.1^{d},10^{e})$	0.13	0.14	0.70	0.66
	4	Poly	$(1^{c}, 0.1^{d},1^{e})$	0.26	0.31	0.28	0.23
	5		$(1^{c}, 0.3^{d},1^{e})$	0.22	0.25	0.46	0.35
	6		$(1^{c},0.1^{d},10^{e})$	0.25	0.30	0.30	0.19
MLR		–	–	0.16	0.23	0.60	0.56
XGBoost	1	n_estimators	$(50^{f}, ^{g})$	0.09	0.10	0.82	0.79
	2		$(100^{f}, ^{g})$	0.13	0.14	0.70	0.66
	3		$(500^{f,g})$	0.02	0.03	0.99	0.96
	4		$1000^{f,g})$	0.12	0.15	0.90	0.70

a: trees, b: randomised state, c: C, d: ε, e: γ, f: estimators, g: maximum depth and learning rate of 4 and 0.1, respectively.

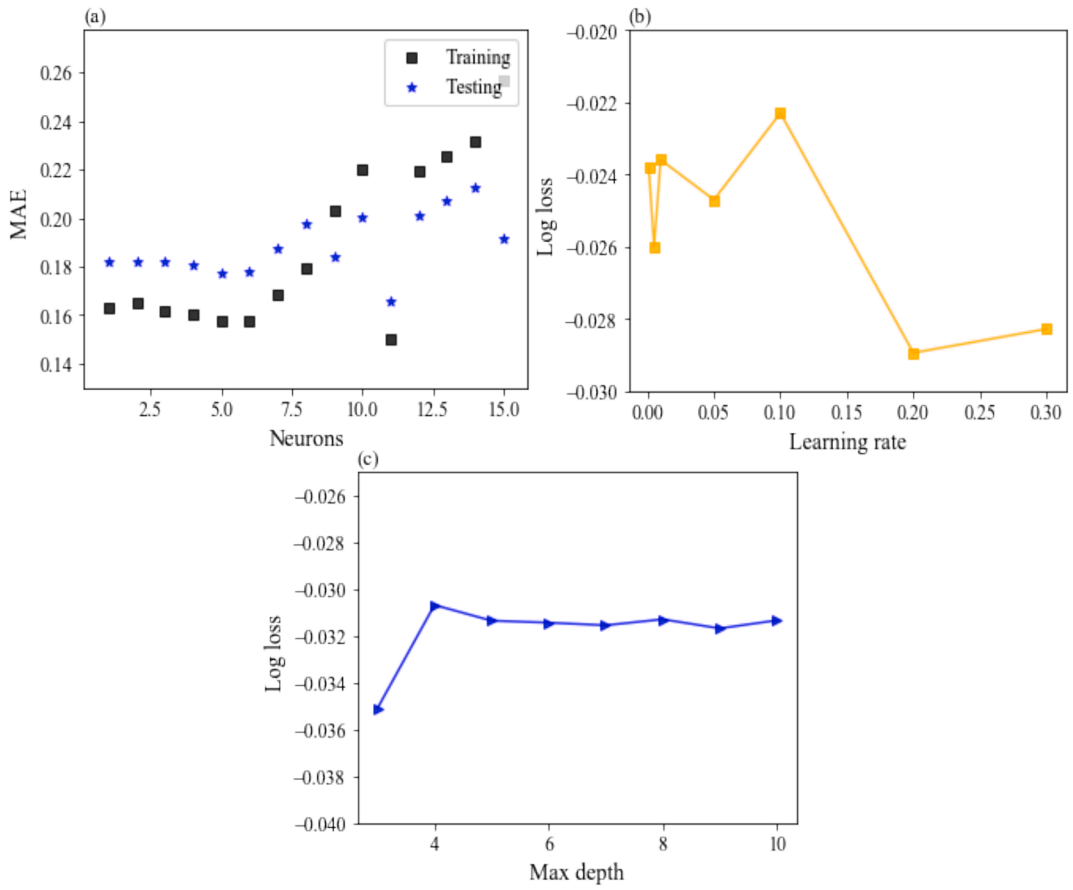

Fig. 5. (a) Number of neurons, (b) learning rates, and (c) maximum depth.

and Musee (2023), where the tanh function performed well. Thus, hyper-parameters are data-dependent and not based on intuition. Moreover, SGD models (ANN 1–3) showed the lowest performance. This feature can be attributed to the position of the gradient vector or the learning processes that are static for all weights (Zarra et al., 2019). In addition, the lowest performance for sigmoid functions was based on the narrow range that does not rescind gradients when saturated and outputs that are not zero-centered (Pushpa and Manimala, 2014). On the other end, the highest performance using the ReLU can be attributed to non-saturation in the positive regime and converging much faster (Rynkiewicz, 2019).

3.3.2. *Random forest regression*

To improve the accuracy of RF models, it is common to use a large number of trees (typically between 100–1000 or higher) (Hou et al., 2020; Liaw and Wiener, 2002). However, in Table 3 the RFR with 100 trees had the highest performance with R = 0.92, and RMSE = 0.06. The models performed admirably when the number of trees increased from 20 to 100 and a further increase from 200 to 500 trees resulted in poor performance. This is because increasing trees beyond the threshold of saturation does not inherently enhance accuracy, especially for small numerical data sets (Oshiro et al., 2012).

3.3.3. *Support vector regression*

In Table 3 the SVR2 was the best model with R and RMSE of 0.87 and 0.10, respectively. Altering the ε from 0.1 (SVR1) to 0.3 (SVR2) improved the accuracy of the models. However, an increase in γ values led to a reduction in R. Low values of γ indicate a large similarity radius, and for high values of γ, the contrast holds. Models with very large γ values tend to overfit (Gretton et al., 2012; Smola and Schölkopf, 2004). The RBF models performed better than Poly. This was in agreement with other previously published results (Papa et al., 2015; Subramanian and Natarajan, 2021; Yalezo and Musee, 2023).

3.3.4. *Extreme gradient boosting*

The XGBoost model is based on the concept of ensemble learning similar to that of RF. Results in Fig. 5b and 5c showed the XGBoost models achieved the highest performance with a learning rate of 0.10 and a maximum depth of 4, respectively. A higher number of estimators lead to better performance and reduce the impact of overfitting as the result of increased diversity and robustness (Osman et al., 2021). However, in Table 3 the 500 estimators were identified as the best model with R of 0.96 and RMSE of 0.03, whereas a further increase to 1000 resulted in over-fitting.

3.4. *Comparison of the performance of ML models*

ANN is regarded as the most extensively used approach for non-trivial problems because of deep learning (Li et al., 2022). However, XGBoost, RFR, and SVR fared better in this investigation. In Fig. 6a and 6b, both XGBoost and RFR had excellent performance in predicting the concentration of Zn^{2+} with R^2 of 0.92 and 0.85, respectively, and low RMSE values. Furthermore, the SVR and ANN models in Fig. 6c and 6d, respectively had a good performance with R^2 in the range of 0.67–0.75. MLR in Fig. 5f yielded the lowest performance with a low R^2 of 0.31 and a large RMSE of 0.23. The results of the MLR algorithms showed ineffectiveness in predicting the concentration of Zn^{2+}.

Violin plots (VP) in Fig. 7 were used to provide a visualisation of the data distribution and to validate these findings. According to an analysis of the distribution shapes in Fig. 7, the XGBoost and RF had matching distribution values at the extreme ends and interquartile range (IQR) to actual data. The SDs of predicted y-values using XGBoost and RF were

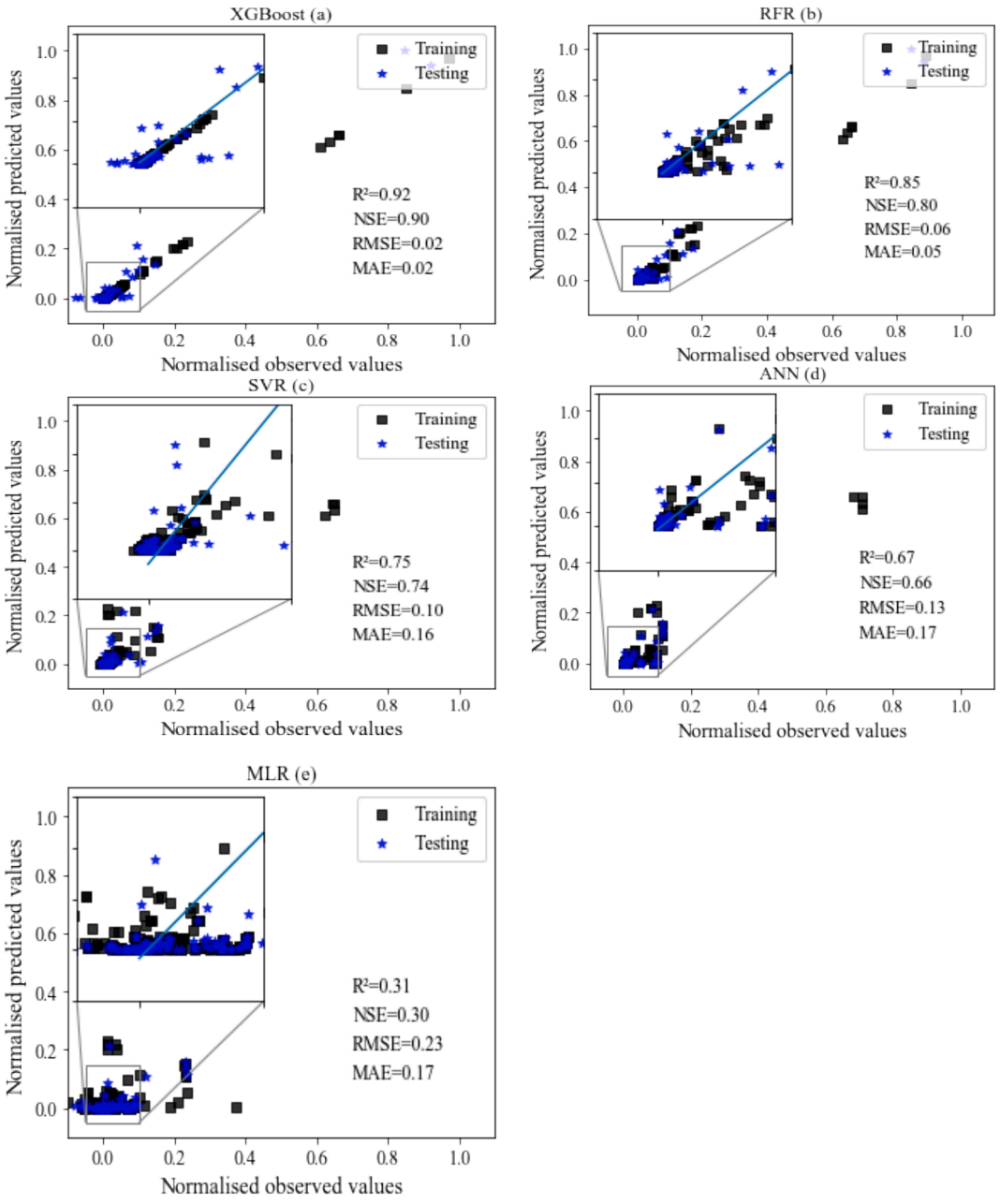

Fig. 6. Scatter plots of the predicted models derived for the dissolution of the nZnO data using NOM, time, nZnO concentration, size, IS, and pH to the concentration of Zn^{2+}. (a) XGBoost, (b) RFR, (c) SVR, (d) ANN, and (e) MLR. Regression line with R^2, NSE, RMSE and MAE metric values.

0.096 and 0.095, respectively; close to the SD of the actual data, which was 0.101. This confirmed the high prediction reliability of both XGBoost and RF. In addition, SDs for SVR and ANN were 0.057 and 0.051, respectively suggesting that these models underestimated the true values. However, MLR exhibited a high degree of overfitting and a large margin of error.

The higher performance by XGBoost was attributed to the approach's optimisation of an arbitrary differentiable loss function using regularisation which reduces overfitting and bias (Dong et al., 2022; Osman et al., 2021). In addition, the RFR model is good error-tolerant, non-parametric, handling non-linearity and lack of data (Wang et al., 2019). On the other hand, SVR generates the output using global minimum; and as such, has a high ability to integrate uncertainty as opposed for example, to ANN where the generated output is based on local minimum (Choubin et al., 2018; Zarei et al., 2018). The poor performance by MLRs was because this modelling approach is based on predefined basic relationships between predictors and output variables, in which, in the circumstance where the data have no fundamental correlation, as in the case of ENP nanoecotoxicology data, the model produce unsatisfactory results (Chen et al., 2019; Zhang et al., 2018).

The results of the ML algorithms in Figs. 6 and 7 in this study have demonstrated the potential to support cost-effective determination and screening of dissolution and, in turn, to reduce cost concomitant with the undertaking of experimental tests for each variation of ENPs using various aquatic permutations (Concu et al., 2017; Furxhi et al., 2019). According to these ML results the parameters of size, pH, IS, NOM, time, and nZnO concentration were identified as reliable prominent variables to screen the dissolution of ENPs in aqueous systems. To account for the individual effect of these parameters mechanistically, the size and surface area of ENPs exhibit an inverse relationship. Smaller nanoparticles dissolve more quickly as particle size decreases (Bian et al., 2011; Domingos et al., 2013).

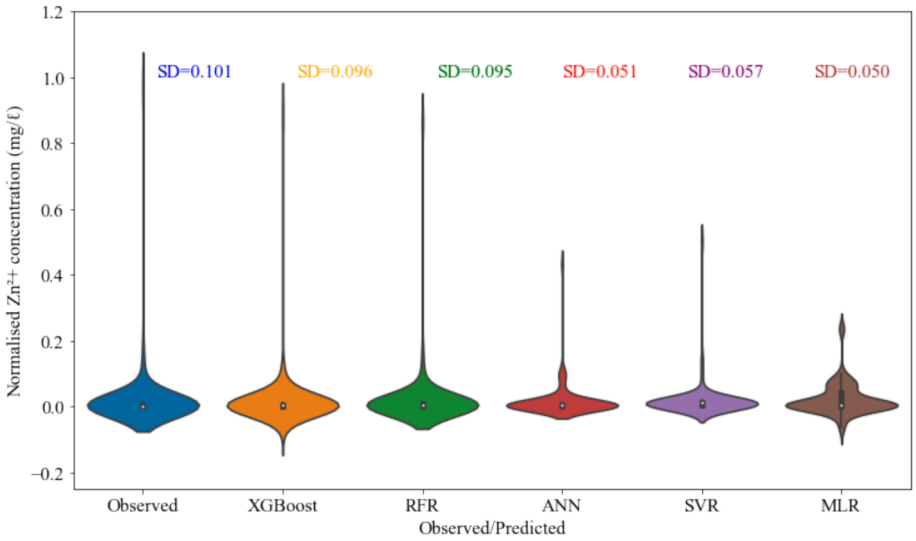

Fig. 7. Visualisation of density mass distribution of the predicted values compared to the observed values based on violin plots (VP) (n = 237). White dots on VP represent the mean of each dataset. The Skinner regime in VP indicates low probability distribution, and contrast holds for the wide regime. The boxes bound indicate the interquartile range (IQR) of the (25th, 50th, and 75th quartiles).

Metal oxide ENPs, such as nZnO, are amphoteric. They undergo dissolution at both acidic and basic pH values. nZnO dissolves more rapidly at low pH values (pH < 6.5) and slower at high pH values (pH ~ 9), which are within the point of zero charges (PZC) region (Han et al., 2016; Han et al., 2014). According to the classical Derjaguin, Landau, Verwey and Overbeek (DLVO) theory, PZC is the vicinity where colloidal particles have strong van der Waals (vdW) and weak electrostatic repulsion forces resulting in high aggregation (Lowry et al., 2012; Schaumann et al., 2015). High rates of dissolution at low pH values occur as the M−O bond weakens due to hydrolysis (Han et al., 2016; Han et al., 2014). Under alkaline pH, dissolution occurs primarily attributed to the complexion soluble hydroxide species M (OH) with polydentate (Bian et al., 2011; Jiang et al., 2015).

Furthermore, the dissolved ions in natural water bodies including anions (NO_3^- SO_4^{2-} Cl^-, etc.) or cations (e.g. Ca^{2+}, K^+, Mg^{2+}, etc.), differ greatly according to the biogeochemical region (Cañedo-Argüelles et al., 2016, 2013; Cormier et al., 2013). High IS or salinity in exposure media is concomitant with a reduction in the chemical potential on the surface of ENPs and, in turn, leads to the dissolution of nZnO (Majedi et al., 2014). General divalent ions have greater impacts than monovalent, even though the type of salt had little significance in this work. In various experimental settings, the rate of dissolution is time-dependent (Bian et al., 2011; Odzak et al., 2017).

NOM constitutes numerous complex biological molecules such as sugars, and cellulosic materials as building blocks (Abbas et al., 2020; Louie et al., 2016). Different categories of NOM include humic substances (humic and fulvic acids), polysaccharides (starch, cellulose, alginate), and proteins (fatty and amino acids) based on differences in molecular weights. Surface functional groups, for example, amide, amine, thiols, hydroxyls, and molecular weight (MW) influence the NOM interactions with ENPs (Philippe and Schaumann, 2014). The impact of NOM on the dissolution undoubtedly points to a diverse set of trends and contradictions. NOM can increase the stability of ENPs, therefore, allowing adequate time for the release of ions. A research study by Han et al. (2014) found that when Suwannee River fulvic acid (SRFA) was present, the Zn^{2+} measurements showed a considerable increase in the aqueous system. Similarly, the addition of citric acid (Mudunkotuwa et al., 2012) and humic acid (HA) (Bian et al., 2011) enhanced the dissolution of ZnO. On the contrary, the presence of NOM can lock the oxidation sites, creating a shielding effect that can limit or prevent dissolution (Hedberg et al., 2019).

3.5. *Randomisation test of developed ML models*

ML models can be prone to random generation; therefore, it was essential to ascertain whether the models are capable of fitting the data more effectively than mere random prediction of noise. Using R^2 as the test statistic, the results of the RT are shown in Fig. 8. According to the H_0 stated in Equation 12, the R^2 computed for the observed data was assumed to have the same distribution as the R^2 after the permutation of data. Based on the results in Fig. 8 it was observed that all developed models had p-values greater than α which was arbitrarily chosen at 0.05. As a result, H_0 was not rejected and, therefore, the developed models were confirmed to not have been randomly generated.

3.6. *Challenges of developed ML models*

Furthermore, the application of ML has demonstrated several benefits in this study, including effectiveness in managing data with uncertainties, ambiguities and non-linearity, as well as its high learning capacity, handling tolerance, low computer code, and ease of updating (Glaubitz et al., 2022; Jordan and Mitchell, 2015; Sun and Scanlon, 2019). However, ML models are typically data-driven and show high reliability and robustness in predicting components that are within the AD range. As a result, the possible limitations of the developed ML models may include the following. First, the ability to adequately generalise and predict the PC and WC properties of ENPs outside of the parameter ranges and regions of high density distribution that are shown in Table S1 and Fig. 9, respectively. In Fig. 9, IS displayed bimodal distribution with high regions of predictability at 0–25 and 70–100 mM. The pH and NOM showed a high distribution between 6.5 and 8.5 and 0–30 mg/ℓ, respectively which represents ranges found in freshwater systems (Abbas et al., 2020; Louie et al., 2016; Troester et al., 2016). Therefore, to refine the model resolution, expansion of the existing ranges by the addition of more distribution points as new data become accessible and the inclusion of other types of ENPs based on class and type not considered in this investigation is necessary.

Second, the use of *meta*-analysis made it possible to gather complete research studies published in the area of interest. However, various studies may not be properly indexed in computer-searchable online databases (Greco et al., 2013; Walker et al., 2008). In addition, data snooping and bias in the original studies can result from merging data using disparate sources through *meta*-analysis (Yalezo and Musee, 2023). Thus, curated nanodatabases such as NanoE-Tox (Juganson et al.,

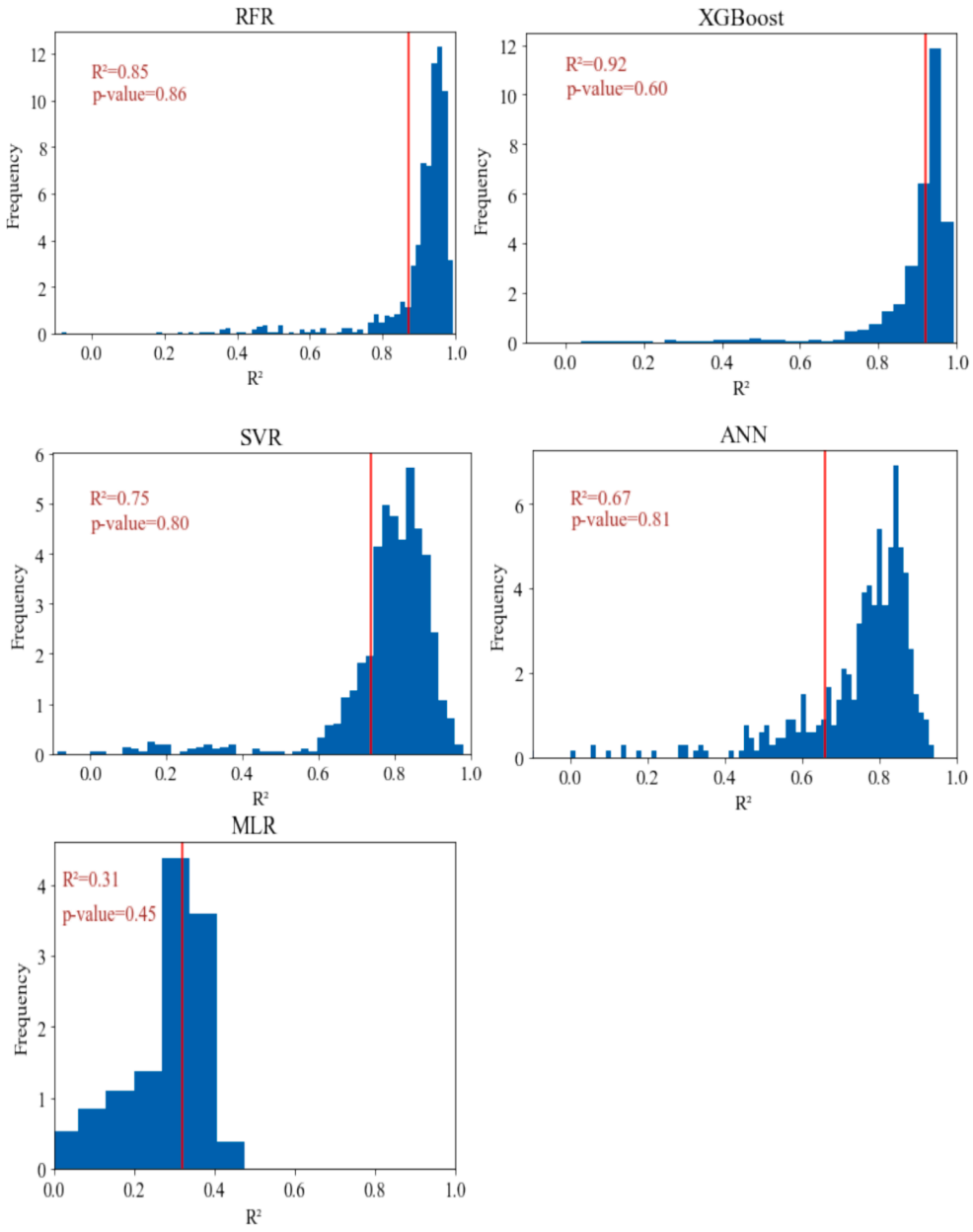

Fig. 8. Results of the randomisation test showing the distribution of permutated results (1000 iterations) against the sampled distribution. R^2 (red line) was used as a test statistic.

2015) and the S2NANO database (www.s2nano.org) (Trinh et al., 2018), must be established, together with standardised experimental protocols, for future investigation.

4. Concluding remarks

Dissolution is a crucial factor influencing the bio persistence and durability of ENPs, which in turn affect their toxicity (Leareng et al., 2020; Mahaye et al., 2017). The current framework for examining the dissolution of ENPs, however, is heavily reliant on experimental testing, which is characterised by ambiguity. This leads to contradictions and a lack of knowledge about the importance of features that affect the transformation processes in aquatic environments for making decisions. Alternatively, the results from this work have demonstrated the suitability of ML tools for initial screening and monitoring nZnO dissolution. This, in turn, guide future experimental investigations by narrowing the focus from many predictors that are concomitant with complexity to an identified smaller number of variables. Our research revealed that continuous input variables such as NOM, time, nZnO concentration, size, IS, and pH are predominant and can be suitable for initial screening and monitoring of the Zn^{2+}concentration in aqueous environment of Among the developed ML models, XGBoost and RFR algorithms were found to be the most effective ML techniques.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used Grammarly and QuillBot to improve language and grammar. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

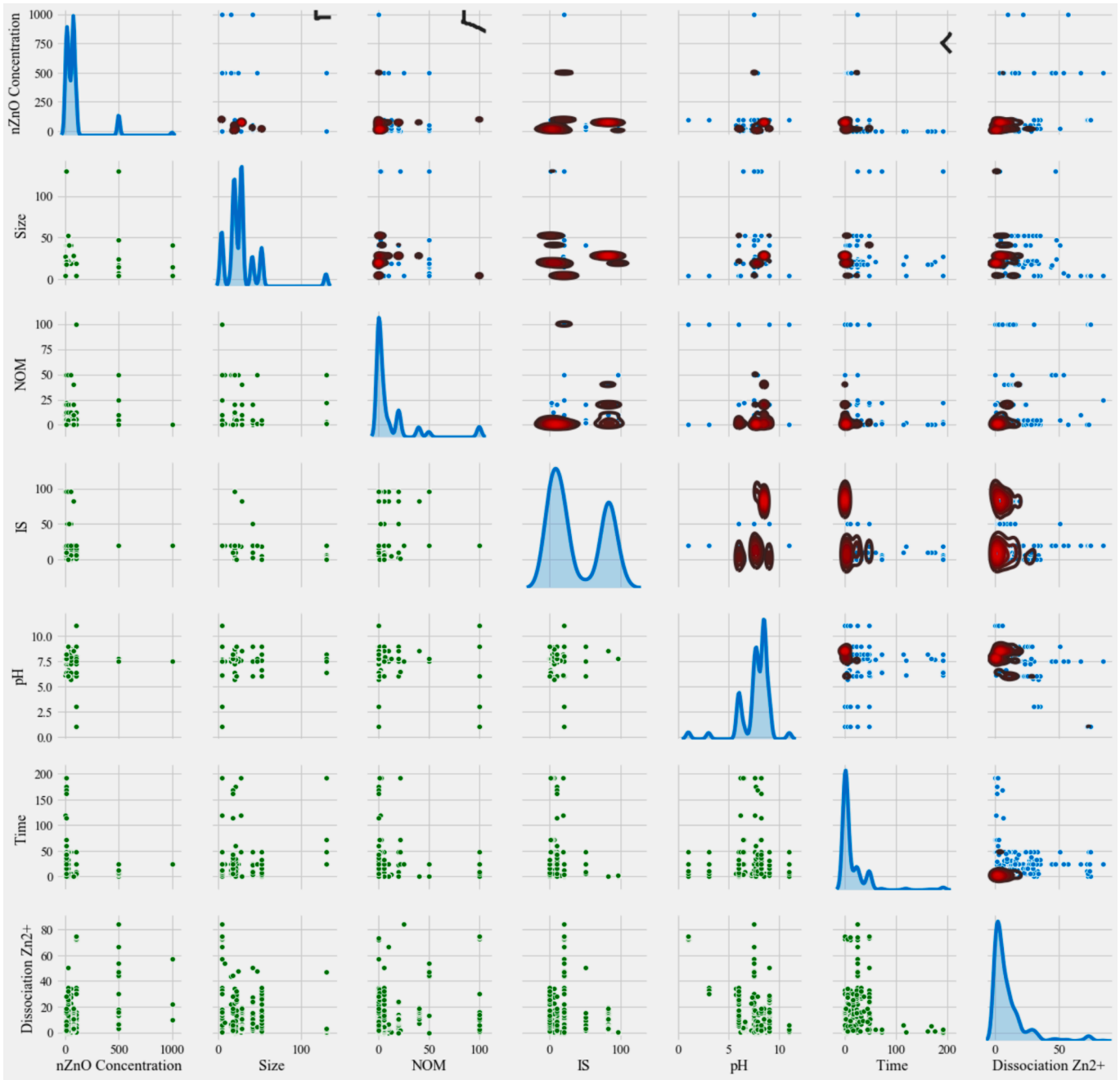

Fig. 9. Pair plots showing the density distribution of input and output parameters in training data to characterised AD. The red circle shows a higher distribution.

CRediT authorship contribution statement

Ntsikelelo Yalezo: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Ndeke Musee:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Michael O. Daramola:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This material is made possible by financial support from the University of Pretoria Grant No. A0Y229, the Water Research Commission Grant No. K5/2509/1, and the National Research Foundation Grant Nos. 112623 and 121170, South Africa.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.enmm.2024.101000.

References

Abbas, Q., Yousaf, B., Ali, M.U., Munir, M.A.M., El-Naggar, A., Rinklebe, J., Naushad, M., 2020. Transformation pathways and fate of engineered nanoparticles (ENPs) in distinct interactive environmental compartments: a review. Environ. Int. 138, 105646.

Alexander, D.L.J., Tropsha, A., Winkler, D.A., 2015. Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J. Chem. Inf. Model. 55, 1316–1322. https://doi.org/10.1021/acs.jcim.5b00206.

Alimissis, A., Philippopoulos, K., Tzanis, C.G., Deligiorgi, D., 2018. Spatial estimation of urban air pollution with the use of artificial neural network models. Atmos. Environ. 191, 205–213. https://doi.org/10.1016/j.atmosenv.2018.07.058.

Bahl, A., Hellack, B., Balas, M., Dinischiotu, A., Wiemann, M., Brinkmann, J., Luch, A., Renard, B.Y., Haase, A., 2019. Recursive feature elimination in random forest classification supports nanomaterial grouping. NanoImpact 15, 100179.

Balraadjsing, S., Peijnenburg, W.J.G.M., Vijver, M.G., 2022. Exploring the potential of in silico machine learning tools for the prediction of acute Daphnia magna nanotoxicity. Chemosphere 307, 135930. https://doi.org/10.1016/j.chemosphere.2022.135930.

Ban, Z., Zhou, Q., Sun, A., Mu, L., Hu, X., 2018. Screening priority factors determining and predicting the reproductive toxicity of various nanoparticles. Environ. Sci. Technol. 52, 9666–9676. https://doi.org/10.1021/acs.est.8b02757.

Ban, Z., Yuan, P., Yu, F., Peng, T., Zhou, Q., Hu, X., 2020. Machine learning predicts the functional composition of the protein corona and the cellular recognition of nanoparticles. Proc. Natl. Acad. Sci. 117, 10492–10499. https://doi.org/10.1073/pnas.1919755117.

Basei, G., Hristozov, D., Lamon, L., Zabeo, A., Jeliazkova, N., Tsiliki, G., Marcomini, A., Torsello, A., 2019. Making use of available and emerging data to predict the hazards of engineered nanomaterials by means of in silico tools: a critical review. NanoImpact 13, 76–99. https://doi.org/10.1016/j.impact.2019.01.003.

Batista, G.E., Monard, M.C., 2003. An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. 17, 519–533.

Bian, S.-W., Mudunkotuwa, I.A., Rupasinghe, T., Grassian, V.H., 2011. Aggregation and dissolution of 4 nm ZnO nanoparticles in aqueous environments: influence of pH, ionic strength, size, and adsorption of humic acid. Langmuir 27, 6059–6068. https://doi.org/10.1021/la200570n.

Cañedo-Argüelles, M., Kefford, B.J., Piscart, C., Prat, N., Schäfer, R.B., Schulz, C.-J., 2013. Salinisation of rivers: an urgent ecological issue. Environ. Pollut. 173, 157–167.

Cañedo-Argüelles, M., Hawkins, C.P., Kefford, B.J., Schäfer, R.B., Dyack, B.J., Brucet, S., Buchwalter, D., Dunlop, J., Frör, O., Lazorchak, J., 2016. Saving freshwater from salts. Science 351, 914–916.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., Janssen, N.A.H., Martin, R.V., Samoli, E., Schwartz, P.E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Vermeulen, R., Brunekreef, B., Hoek, G., 2019. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int. 130, 104934 https://doi.org/10.1016/j.envint.2019.104934.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., 2015. Xgboost: extreme gradient boosting. R Package Version 04–2 (1), 1–4.

Chen, X., O'Halloran, J., Jansen, M.A.K., 2016. The toxicity of zinc oxide nanoparticles to Lemna minor (L.) is predominantly caused by dissolved Zn. Aquat. Toxicol. 174, 46–53. https://doi.org/10.1016/j.aquatox.2016.02.012.

Choi, J.-S., Ha, M.K., Trinh, T.X., Yoon, T.H., Byun, H.-G., 2018. Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources. Sci. Rep. 8, 6110. https://doi.org/10.1038/s41598-018-24483-z.

Choubin, B., Darabi, H., Rahmati, O., Sajedi-Hosseini, F., Kløve, B., 2018. River suspended sediment modelling using the CART model: a comparative study of machine learning techniques. Sci. Total Environ. 615, 272–281. https://doi.org/10.1016/j.scitotenv.2017.09.293.

Chowdhury, I., Cwiertny, D.M., Walker, S.L., 2012. Combined factors influencing the aggregation and deposition of nano-TiO_2 in the presence of humic acid and bacteria. Environ. Sci. Technol. 46, 6968–6976. https://doi.org/10.1021/es2034747.

Ciszewski, M.G., Söhl, J., Leenen, T., Van Trigt, B., Jongbloed, G., 2024. Testing for no effect in regression problems: a permutation approach. Stat. Neerlandica stan.12346. https://doi.org/10.1111/stan.12346.

Concu, R., Kleandrova, V.V., Speck-Planche, A., Cordeiro, M.N.D.S., 2017. Probing the toxicity of nanoparticles: a unified *in silico* machine learning model based on perturbation theory. Nanotoxicology 11, 891–906. https://doi.org/10.1080/17435390.2017.1379567.

Cormier, S.M., Suter, G.W., Zheng, L., 2013. Derivation of a benchmark for freshwater ionic strength. Environ. Toxicol. Chem. 32, 263–271.

Debanath, M.K., Karmakar, S., 2013. Study of blueshift of optical band gap in zinc oxide (ZnO) nanoparticles prepared by low-temperature wet chemical method. Mater. Lett. 111, 116–119.

Deji, Z., Liu, P., Wang, X., Zhang, X., Luo, Y., Huang, Z., 2021. Association between maternal exposure to perfluoroalkyl and polyfluoroalkyl substances and risks of adverse pregnancy outcomes: a systematic review and meta-analysis. Sci. Total Environ. 783, 146984 https://doi.org/10.1016/j.scitotenv.2021.146984.

Domingos, R.F., Rafiei, Z., Monteiro, C.E., Khan, M.A.K., Wilkinson, K.J., 2013. Agglomeration and dissolution of zinc oxide nanoparticles: role of pH, ionic strength and fulvic acid. Environ. Chem. 10, 306. https://doi.org/10.1071/EN12202.

Dong, S., Wu, Z., Wang, M., Sun, X., Mao, L., 2022. Assessing comparable bioconcentration potentials for nanoparticles in aquatic organisms via combined utilization of machine learning and toxicokinetic models. SmartMat smm2.1155. https://doi.org/10.1002/smm2.1155.

Findlay, M.R., Freitas, D.N., Mobed-Miremadi, M., Wheeler, K.E., 2018. Machine learning provides predictive analysis into silver nanoparticle protein corona formation from physicochemical properties. Environ. Sci. Nano 5, 64–71. https://doi.org/10.1039/C7EN00466D.

Fjodorova, N., Novic, M., Gajewicz, A., Rasulev, B., 2017. The way to cover prediction for cytotoxicity for all existing nano-sized metal oxides by using neural network method. Nanotoxicology 11, 475–483. https://doi.org/10.1080/17435390.2017.1310949.

Foley, C.J., Feiner, Z.S., Malinich, T.D., Höök, T.O., 2018. A meta-analysis of the effects of exposure to microplastics on fish and aquatic invertebrates. Sci. Total Environ. 631, 550–559.

Foss Hansen, S., Heggelund, L.R., Revilla Besora, P., Mackevica, A., Boldrin, A., Baun, A., 2016. Nanoproducts – what is actually available to European consumers? Environ. Sci. Nano 3, 169–180. https://doi.org/10.1039/C5EN00182J.

Furxhi, I., Murphy, F., Mullins, M., Poland, C.A., 2019. Machine learning prediction of nanoparticle in vitro toxicity: a comparative study of classifiers and ensemble-classifiers using the Copeland Index. Toxicol. Lett. 312, 157–166. https://doi.org/10.1016/j.toxlet.2019.05.016.

Gagliardi, B.S., Pettigrove, V.J., Long, S.M., Hoffmann, A.A., 2016. A meta-analysis evaluating the relationship between aquatic contaminants and chironomid larval deformities in laboratory studies. Environ. Sci. Technol. 50, 12903–12911. https://doi.org/10.1021/acs.est.6b04020.

Glaubitz, C., Rothen-Rutishauser, B., Lattuada, M., Balog, S., Petri-Fink, A., 2022. Designing the ultrasonic treatment of nanoparticle-dispersions *via* machine learning. Nanoscale 14, 12940–12950. https://doi.org/10.1039/D2NR03240F.

Goldberg, E., Scheringer, M., Bucheli, T.D., Hungerbühler, K., 2015. Prediction of nanoparticle transport behavior from physicochemical properties: machine learning provides insights to guide the next generation of transport models. Environ. Sci. Nano 2, 352–360. https://doi.org/10.1039/C5EN00050E.

Greco, T., Zangrillo, A., Biondi-Zoccai, G., Landoni, G., 2013. Meta-analysis: pitfalls and hints. Heart Lung Vessels 5, 219.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012. A kernel two-sample test. J. Mach. Learn. Res. 13, 723–773.

Grillo, R., De Jesus, M.B., Fraceto, L.F., 2018. Environmental impact of nanotechnology: analyzing the present for building the future. Front. Environ. Sci.

Gurevitch, J., Koricheva, J., Nakagawa, S., Stewart, G., 2018. Meta-analysis and the science of research synthesis. Nature 555, 175–182.

Han, Y., Kim, D., Hwang, G., Lee, B., Eom, I., Kim, P.J., Tong, M., Kim, H., 2014. Aggregation and dissolution of ZnO nanoparticles synthesized by different methods: influence of ionic strength and humic acid. Colloids Surf. Physicochem. Eng. Asp. 451, 7–15. https://doi.org/10.1016/j.colsurfa.2014.03.030.

Han, Y., Hwang, G., Kim, D., Bradford, S.A., Lee, B., Eom, I., Kim, P.J., Choi, S.Q., Kim, H., 2016. Transport, retention, and long-term release behavior of ZnO nanoparticle aggregates in saturated quartz sand: Role of solution pH and biofilm coating. Water Res. 90, 247–257. https://doi.org/10.1016/j.watres.2015.12.009.

Hanser, T., Barber, C., Marchaland, J.F., Werner, S., 2016. Applicability domain: towards a more formal definition. SAR QSAR Environ. Res. 27, 865–881. https://doi.org/10.1080/1062936X.2016.1250229.

Hedberg, J., Blomberg, E., Odnevall Wallinder, I., 2019. In the search for nanospecific effects of dissolution of metallic nanoparticles at freshwater-like conditions: a critical review. Environ. Sci. Technol. 53, 4030–4044.

Hou, P., Jolliet, O., Zhu, J., Xu, M., 2020. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. Environ. Int. 135, 105393 https://doi.org/10.1016/j.envint.2019.105393.

Hou, J., Wu, Y., Li, X., Wei, B., Li, S., Wang, X., 2018. Toxic effects of different types of zinc oxide nanoparticles on algae, plants, invertebrates, vertebrates and microorganisms. Chemosphere 193, 852–860. https://doi.org/10.1016/j.chemosphere.2017.11.077.

Jiang, C., Aiken, G.R., Hsu-Kim, H., 2015. Effects of natural organic matter properties on the dissolution kinetics of zinc oxide nanoparticles. Environ. Sci. Technol. 49, 11476–11484. https://doi.org/10.1021/acs.est.5b02406.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349, 255–260.

Juganson, K., Ivask, A., Blinova, I., Mortimer, M., Kahru, A., 2015. NanoE-Tox: new and in-depth database concerning ecotoxicity of nanomaterials. Beilstein J. Nanotechnol. 6, 1788–1804. https://doi.org/10.3762/bjnano.6.183.

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., Vermeulen, R.C.H., 2019. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. Environ. Sci. Technol. 53, 1413–1421. https://doi.org/10.1021/acs.est.8b06038.

Khoshroo, A., Emrouznejad, A., Ghaffarizadeh, A., Kasraei, M., Omid, M., 2018. Sensitivity analysis of energy inputs in crop production using artificial neural networks. J. Clean. Prod. 197, 992–998. https://doi.org/10.1016/j.jclepro.2018.05.249.

Leareng, S.K., Ubomba-Jaswa, E., Musee, N., 2020. Toxicity of zinc oxide and iron oxide engineered nanoparticles to *Bacillus subtilis* in river water systems. Environ. Sci. Nano 7, 172–185. https://doi.org/10.1039/C9EN00585D.

Lee, J., Im, J., Kim, U., Löffler, F.E., 2016. A Data mining approach to predict in situ detoxification potential of chlorinated ethenes. Environ. Sci. Technol. 50, 5181–5188. https://doi.org/10.1021/acs.est.5b05090.

Li, J., Wang, C., Yue, L., Chen, F., Cao, X., Wang, Z., 2022. Nano-QSAR modeling for predicting the cytotoxicity of metallic and metal oxide nanoparticles: a review. Ecotoxicol. Environ. Saf. 243, 113955 https://doi.org/10.1016/j.ecoenv.2022.113955.

Li, M., Zhu, L., Lin, D., 2011. Toxicity of ZnO nanoparticles to *Escherichia coli*: mechanism and the influence of medium components. Environ. Sci. Technol. 45, 1977–1983. https://doi.org/10.1021/es102624t.
Liaw, A., Wiener, M., 2002. Classification and regression by random. Forest 2, 6.
Lo, A., Chernoff, H., Zheng, T., Lo, S.-H., 2015. Why significant variables aren't automatically good predictors. Proc. Natl. Acad. Sci. 112, 13892–13897. https://doi.org/10.1073/pnas.1518285112.
Lodeiro, P., Achterberg, E.P., Pampín, J., Affatati, A., El-Shahawi, M.S., 2016. Silver nanoparticles coated with natural polysaccharides as models to study AgNP aggregation kinetics using UV-Visible spectrophotometry upon discharge in complex environments. Sci. Total Environ. 539, 7–16. https://doi.org/10.1016/j.scitotenv.2015.08.115.
Louie, S.M., Tilton, R.D., Lowry, G.V., 2013. Effects of molecular weight distribution and chemical properties of natural organic matter on gold nanoparticle aggregation. Environ. Sci. Technol. 47, 4245–4254. https://doi.org/10.1021/es400137x.
Louie, S.M., Tilton, R.D., Lowry, G.V., 2016. Critical review: impacts of macromolecular coatings on critical physicochemical processes controlling environmental fate of nanomaterials. Environ. Sci. Nano 3, 283–310. https://doi.org/10.1039/C5EN00104H.
Lowry, G.V., Gregory, K.B., Apte, S.C., Lead, J.R., 2012. Transformations of nanomaterials in the environment. Environ. Sci. Technol. 46, 6893–6899. https://doi.org/10.1021/es300839e.
Mahaye, N., Thwala, M., Cowan, D.A., Musee, N., 2017. Genotoxicity of metal based engineered nanoparticles in aquatic organisms: a review. Mutat. Res. Mutat. Res. 773, 134–160. https://doi.org/10.1016/j.mrrev.2017.05.004.
Majedi, S.M., Kelly, B.C., Lee, H.K., 2014. Role of combinatorial environmental factors in the behavior and fate of ZnO nanoparticles in aqueous systems: a multiparametric analysis. J. Hazard. Mater. 264, 370–379.
Mirzaei, M., Furxhi, I., Murphy, F., Mullins, M., 2021. A machine learning tool to predict the antibacterial capacity of nanoparticles. Nanomaterials 11, 1774. https://doi.org/10.3390/nano11071774.
Mudunkotuwa, I.A., Rupasinghe, T., Wu, C.-M., Grassian, V.H., 2012. Dissolution of ZnO nanoparticles at circumneutral pH: a study of size effects in the presence and absence of citric acid. Langmuir 28, 396–403. https://doi.org/10.1021/la203542x.
Musee, N., Zvimba, J.N., Schaefer, L.M., Nota, N., Sikhwivhilu, L.M., Thwala, M., 2014. Fate and behavior of ZnO-and Ag-engineered nanoparticles and a bacterial viability assessment in a simulated wastewater treatment plant. J. Environ. Sci. Health Part A 49, 59–66.
Ni, J., Wu, G.D., Albenberg, L., Tomov, V.T., 2017. Gut microbiota and IBD: causation or correlation? Nat. Rev. Gastroenterol. Hepatol. 14, 573–584.
Odzak, N., Kistler, D., Sigg, L., 2017. Influence of daylight on the fate of silver and zinc oxide nanoparticles in natural aquatic environments. Environ. Pollut. 226, 1–11. https://doi.org/10.1016/j.envpol.2017.04.006.
Ogunleye, A., Wang, Q.-G., 2019. XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans. Comput. Biol. Bioinform. 17, 2131–2140.
Ojala, M., Garriga, G.C., 2010. Permutation tests for studying classifier performance. J. Mach. Learn. Res. 11.
Oshiro, T.M., Perez, P.S., Baranauskas, J.A., 2012. How many trees in a random forest?, in: Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8. Springer, pp. 154–168.
Osman, A.I.A., Ahmed, A.N., Chow, M.F., Huang, Y.F., El-Shafie, A., 2021. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Eng. J. 12, 1545–1556.
Papa, E., Doucet, J.P., Doucet-Panaye, A., 2015. Linear and non-linear modelling of the cytotoxicity of TiO $_2$ and ZnO nanoparticles by empirical descriptors. SAR QSAR Environ. Res. 26, 647–665. https://doi.org/10.1080/1062936X.2015.1080186.
Parihar, V., Raja, M., Paulose, R., 2018. A brief review of structural, electrical and electrochemical properties of zinc oxide nanoparticles. Rev. Adv. Mater. Sci. 53, 119–130. https://doi.org/10.1515/rams-2018-0009.
Peng, T., Wei, C., Yu, F., Xu, J., Zhou, Q., Shi, T., Hu, X., 2020. Predicting nanotoxicity by an integrated machine learning and metabolomics approach. Environ. Pollut. 267, 115434 https://doi.org/10.1016/j.envpol.2020.115434.
Philippe, A., Schaumann, G.E., 2014. Interactions of dissolved organic matter with natural and engineered inorganic colloids: a review. Environ. Sci. Technol. 48, 8946–8962. https://doi.org/10.1021/es502342r.
Pushpa, P., Manimala, K., 2014. Implementation of hyperbolic tangent activation function in VLSI. Int. J. Adv. Res. Comput. Sci. Technol. 2, 225–228.
Rynkiewicz, J., 2019. Asymptotic statistics for multilayer perceptron with ReLU hidden units. Neurocomputing 342, 16–23. https://doi.org/10.1016/j.neucom.2018.11.097.
Schaumann, G.E., Philippe, A., Bundschuh, M., Metreveli, G., Klitzke, S., Rakcheev, D., Grün, A., Kumahor, S.K., Kühn, M., Baumann, T., Lang, F., Manz, W., Schulz, R., Vogel, H.-J., 2015. Understanding the fate and biological effects of Ag- and TiO2-nanoparticles in the environment: The quest for advanced analytics and interdisciplinary concepts. Sci. Total Environ. 535, 3–19. https://doi.org/10.1016/j.scitotenv.2014.10.035.
Schiavo, S., Oliviero, M., Miglietta, M., Rametta, G., Manzo, S., 2016. Genotoxic and cytotoxic effects of ZnO nanoparticles for Dunaliella tertiolecta and comparison with SiO 2 and TiO 2 effects at population growth inhibition levels. Sci. Total Environ. 550, 619–627. https://doi.org/10.1016/j.scitotenv.2016.01.135.
Sengul, A.B., Asmatulu, E., 2020. Toxicity of metal and metal oxide nanoparticles: a review. Environ. Chem. Lett. 18, 1659–1683.
Sharma, V.K., Siskova, K.M., Zboril, R., Gardea-Torresdey, J.L., 2014. Organic-coated silver nanoparticles in biological and environmental conditions: fate, stability and toxicity. Adv. Colloid Interface Sci. 204, 15–34. https://doi.org/10.1016/j.cis.2013.12.002.
Shipley, B., 2016. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R. Cambridge University Press.
Siepmann, J., Siepmann, F., 2013. Mathematical modeling of drug dissolution. Int. J. Pharm. 453, 12–24.
Sirelkhatim, A., Mahmud, S., Seeni, A., Kaus, N.H.M., Ann, L.C., Bakhori, S.K.M., Hasan, H., Mohamad, D., 2015. Review on zinc oxide nanoparticles: antibacterial activity and toxicity mechanism. Nano-Micro Lett. 7, 219–242. https://doi.org/10.1007/s40820-015-0040-x.
Sizochenko, N., Syzochenko, M., Fjodorova, N., Rasulev, B., Leszczynski, J., 2019. Evaluating genotoxicity of metal oxide nanoparticles: Application of advanced supervised and unsupervised machine learning techniques. Ecotoxicol. Environ. Saf. 185, 109733 https://doi.org/10.1016/j.ecoenv.2019.109733.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222.
Song, Y., Rottschäfer, V., Vijver, M.G., Peijnenburg, W.J., 2023. Developing and verifying a quantitative dissolution model for metal-bearing nanoparticles in aqueous media. Environ. Sci. Nano 10, 1790–1799.
Subramanian, D., Natarajan, J., 2021. Integrated meta-analysis and machine learning approach identifies acyl-CoA thioesterase with other novel genes responsible for biofilm development in Staphylococcus aureus. Infect. Genet. Evol. 88, 104702 https://doi.org/10.1016/j.meegid.2020.104702.
Subramanian, N.A., Palaniappan, A., 2021. NanoTox: development of a parsimonious *in silico* model for toxicity assessment of metal-oxide nanoparticles using physicochemical features. ACS Omega 6, 11729–11739. https://doi.org/10.1021/acsomega.1c01076.
Sun, A.Y., Scanlon, B.R., 2019. How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environ. Res. Lett. 14, 073001 https://doi.org/10.1088/1748-9326/ab1b7d.
Takahashi, K., Takahashi, L., 2019. Data driven determination in growth of silver from clusters to nanoparticles and bulk. J. Phys. Chem. Lett. 10, 4063–4068. https://doi.org/10.1021/acs.jpclett.9b01394.
Trinh, T.X., Ha, M.K., Choi, J.S., Byun, H.G., Yoon, T.H., 2018. Curation of datasets, assessment of their quality and completeness, and nanoSAR classification model development for metallic nanoparticles. Environ. Sci. Nano 5, 1902–1910.
Troester, M., Brauch, H.-J., Hofmann, T., 2016. Vulnerability of drinking water supplies to engineered nanoparticles. Water Res. 96, 255–279.
Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525. https://doi.org/10.1093/bioinformatics/17.6.520.
Utembe, W., Potgieter, K., Stefaniak, A.B., Gulumian, M., 2015. Dissolution and biodurability: important parameters needed for risk assessment of nanomaterials. Part. Fibre Toxicol. 12, 11. https://doi.org/10.1186/s12989-015-0088-2.
Valente, G., Castellanos, A.L., Hausfeld, L., De Martino, F., Formisano, E., 2021. Cross-validation and permutations in MVPA: validity of permutation strategies and power of cross-validation schemes. Neuroimage 238, 118145.
Walker, E., Hernandez, A.V., Kattan, M.W., 2008. Meta-analysis: its strengths and limitations. Cleve. Clin. J. Med. 75, 431.
Wang, Y., Dong, H., Zhu, Z., Gerber, P.J., Xin, H., Smith, P., Opio, C., Steinfeld, H., Chadwick, D., 2017. Mitigating greenhouse gas and ammonia emissions from swine manure management: a system analysis. Environ. Sci. Technol. 51, 4503–4511. https://doi.org/10.1021/acs.est.6b06430.
Wang, Y., Du, Y., Wang, J., Li, T., 2019. Calibration of a low-cost PM2.5 monitor using a random forest model. Environ. Int. 133, 105161 https://doi.org/10.1016/j.envint.2019.105161.
Wang, P., Meng, P., Zhai, J.-Y., Zhu, Z.-Q., 2013. A hybrid method using experiment design and grey relational analysis for multiple criteria decision making problems. Knowl.-Based Syst. 53, 100–107. https://doi.org/10.1016/j.knosys.2013.08.025.
Yalezo, N., Musee, N., 2023. Meta-analysis of engineered nanoparticles dynamic aggregation in freshwater-like systems using machine learning techniques. J. Environ. Manage. 337, 117739.
Zarei, T., Behyad, R., Abedini, E., 2018. Study on parameters effective on the performance of a humidification-dehumidification seawater greenhouse using support vector regression. Desalination 435, 235–245. https://doi.org/10.1016/j.desal.2017.05.033.
Zarra, T., Galang, M.G., Ballesteros, F., Belgiorno, V., Naddeo, V., 2019. Environmental odour management by artificial neural network – a review. Environ. Int. 133, 105189 https://doi.org/10.1016/j.envint.2019.105189.
Zhang, Y., Chen, H., Yang, B., Fu, S., Yu, J., Wang, Z., 2018. Prediction of phosphate concentrate grade based on artificial neural network modeling. Results Phys. 11, 625–628. https://doi.org/10.1016/j.rinp.2018.10.011.
Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R., 2017. Efficient kNN classification with different numbers of nearest neighbors. IEEE Trans. Neural Netw. Learn. Syst. 29, 1774–1785.