UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Application of Machine Learning and Knowledge-Based Systems to Support Decision-Making on Fate and Behaviour of Engineered Nanoparticles in Aqueous Environment

By

Ntsikelelo Yalezo

(10412728)

A thesis submitted in partial fulfillment of the requirements for the degree

Doctor of Philosophy

In the

Department of Chemical Engineering,

Faculty of Engineering, the Built Environment, and Information Technology

University of Pretoria

Supervisors:

Prof Michael O Daramola and Prof Ndeke Musee

January 2025

Declaration by candidate

I hereby declare that this is a record of my original work carried out in fulfilment of the requirements for the degree of PhD in the University of Pretoria and has not previously been submitted to any other institution of higher learning. I further declare that all sources cited or quoted are indicated and acknowledged by a comprehensive list of references.

Candidate

Acknowledgements

- I'm grateful that God Almighty gave me life. I express heartfelt appreciation to both Professor Michael Daramola and Professor Ndeke Musee for their support during this journey. It was a pleasure to work under their guidance.

- To the University of Pretoria's Department of Chemical Engineering for hosting the project. The Water Research Commission and the National Research Foundation for funding. I would like to express my gratitude to my former research colleagues in the Emerging Contaminants Ecological and Risk Assessment (ECERA) group. I appreciate your thoughtful observations and constructive criticism. Dr. Samuel Leareng, Dr. Ntombikayise Mahaye, Dr. Aston Nanja, Cornelius Marthinus van der Walt, Ernst Bekker, and Mpho Makofane. Additionally, special thanks to Dr Leareng and Nanja for proofreading this thesis

- In closing, I would want to thank my family members: Father; Bhutina Yalezo, Uncle: Mr Thembile Ndzungeni, Siblings: Mr. Vuyo Yalezo, Mr. Sabelo Yalezo, Sinazo Pawaula, Fiance; Khanyiswa Cubeni, Children; Khwezi, Asemahle, and Iminathi, for their love and unwavering support.

Dedication

This study is dedicated to my late mother

Buyiswa Cynthia "mamnqarhwane" Yalezo,

Lala kakuhle masiduli

List of research outputs

Journals

Published articles

- N. Yalezo and N. Musee. Meta-analysis for aggregation of engineered nanoparticles in freshwater-like systems using machine learning techniques. Journal of Environmental Management 337 (2023): 117739. https://doi.org/10.1016/j.jenvman.2023.117739.
- N. Yalezo, N. Musee and M. O Daramola. Developing machine learning algorithm for predicting the dissolution of zinc oxide nanoparticles in freshwater-like systems. Environmental Nanotechnology, Monitoring & Management, Volume 22, (2024), 101000. https://doi.org/10.1016/j.enmm.2024.101000.

Under review

- N. Yalezo, N. Musee and M. O Daramola. A model for screening the fate and behaviour of the engineered nanoparticles in aquatic systems using semi-quantitative analysis and decision tree classifiers. NanoImpact **IMPACT-D-24-00160**.
- N. Yalezo, N. Musee and M. O Daramola. A model using fuzzy logic for assessing the fate and behaviour of engineered nanoparticles in the aqueous environment. Next Research. **NEXRES-24-00674**.

In preparation

- N. Yalezo, N. Musee and M. O Daramola. Application of machine learning algorithms to advance risk assessment in nanoecotoxicology domain; status quo, challenges and perspective: Review.

Technical reports as Chapters

- Yalezo, Ntsikelelo, and Ndeke Musee. "Modelling the aggregation of engineered nanoparticles in aquatic systems using data-driven techniques." *risk assessment on nano-and macro-scale emerging contaminants in freshwater systems using experimental and modelling techniques (2022):* Volume 1. https://www.wrc.org.za/?mdocs-file=63611#page=41.

- Yalezo, Ntsikelelo, and Ndeke Musee. "Development of decision support system to estimate exposure potential of ENPs in aquatic systems using fuzzy logic theory." *risk assessment on nano-and macro-scale emerging contaminants in freshwater systems using experimental and modelling techniques* (2022): Volume 3, https://www.wrc.org.za/?mdocs-file=63605#page=69.

List of Conference presentations

- **N. Yalezo** and N. Musee. 8th International Conference on Environmental, Health and Safety issues related to Nanomaterials, (NanoSafe 2023), *June 5-9, 2023. Maison Minatec, Grenoble, France.* ***(Oral presentation).***
- **N. Yalezo** and N. Musee. Assessment of metal oxide-based nanoparticles stability in the aquatic systems using fuzzy logic. *6th International Conference on Health and safety issues related to Nanomaterials. (NanoSafe 2018), 5-9 November, MINATEC, Grenoble, France.* ***(Oral presentation).***
- Musee, N., Leareng, S. K., Mahaye, N., Nanja, A. F., Bekker, E. H., **Yalezo, N**., 2018. Trends on environmental transformations and effects of nanomaterials mixtures in aquatic systems: an overview. *6th International Conference on Health and safety issues related to Nanomaterials. (NanoSafe 2018), 5-9 November, MINATEC, Grenoble, France.*

Synopsis

In recent decades, modern science has transformed due to the recognition and usage of engineered nanoparticles (ENP), which are substances that have a peripheral dimension in the geometric range of 1 to 100 nm. These materials have evolved into multifunctional materials that are used for engineering and innovations in a wide range of fields, including agriculture, medicine, food, industry, biomedical, and energy. Apart from their unique functionality and numerous benefits, the surge in production of multi-enabled nano-products and their emission in ecology systems, specifically to aquatic systems has raised serious environmental concerns about the potential deleterious effects of ENPs on the aquatic biota. So far, to address the existing environmental safety concerns, a volume of experimental data has been generated using natural freshwater and like systems for characterisation of the ENP colloidal stability. However, this data is knowledge-poor, heterogeneous, highly multifaceted (not easily discernible data variables relationship), and uncertain (due to the multiplicity of data); thus, highly challenging to be utilised to support the decision-making.

Therefore, this work describes the application of data modelling techniques for the development of computer-based intelligent systems to support decision-making in dealing with the fate and behaviour of ENPs. At the same time, the study aims to provide a coherent understanding of the mechanisms and interactions that underpin these ENP transformation behaviours in aqueous environments. The field of data modelling has gained prominence across various fields including numerous environmental domains with the advancement of artificial intelligence (AI) research, digital computers, and big data. The modelling techniques of interest in this study included machine learning (ML) (i.e., adaptive neuro-fuzzy inference system (ANFIS), artificial neural network (ANN), support vector regression (SVR), random forest regression (RFR), k-nearest neighbour (KNN), extreme gradient boosting (XGBoost) and multi-linear regression (MLR)) and knowledge-based system (KBS) (i.e., fuzzy logic, semi-quantitative analysis).

The results showed that ML was quite useful for modelling heterogeneity and non-linear data. It also revealed that diverse ENP transformation processes are influenced by variant parameters and that significant variables reported experimentally are not

fundamentally good predictive variables. The RFR algorithms had the highest performance with the coefficient of determination (R^2) and Nash-Sutcliffe efficiency (NSE) greater than 0.80 and 0.70, respectively for predicting the dynamic aggregation of ENPs. Furthermore, to predict the dissolution of nZnO the models that performed the best were the RFR and XGBoost algorithms with R^2 values of 0.85 and 0.92, respectively. Overall, ML techniques of RFR, XGBoost, SVR, and ANN models yielded satisfactory to a very good level of accuracy in predicting both the aggregation and dissolution of ENPs. However, MLR showed poor performance, for both processes an indication of no underlying linear relationship between the model inputs and output.

In addition, to ML algorithms demonstrating high prediction accuracy, and meta-analysis aiding to quantitatively evaluate highly heterogeneous data from multiple literature sources: to account for the scarcity of quantitative data the domain knowledge was encoded using rules and scores to develop intelligent KBS. These included computer-based semi-quantitative analysis integrated with decision tree classifiers (SQADTC) and fuzzy decision-making systems (FDMS). SQADTC used several weights/scores allocated to different factors (inputs) or linguistic variables and their sub-level (intermediate outputs). The functionality of SQADTC was illustrated using worked case studies of silver (nAg), nZnO, and $nTiO_2$. The results demonstrate that our proposed model can be highly effective and valuable for preliminary screening of the exposure of ENPs. SQA application is relatively cost-effective and easy to use since no software or computational tools are required. In addition, non-experts can easily understand the hierarchical nature, Boolean logic, and visual representations of DTCs; which is highly valuable given that testing each variation of ENPs is tedious and associated with high cost.

Furthermore, the FDMS constituted 321 (three hundred and twenty-one) if-then conditional statements in the fuzzy inference system. Modelling results using FDMS in the case studies of $nTiO_2$ and nZnO demonstrated that the representation of qualitative knowledge by fuzzy sets and its application as very successful in handling partial truth information. FDMS provides flexibility to reduce bias and integrate the uncertainty that arises with the modelling of expert intuitions or perceptions. FDMS was able to replicate human-like reasoning using natural language in complex

scenarios with no sharp boundaries, which makes the model ideal in various real-world scenarios.

Overall, this thesis work offers the application of ML and KBS as a basis to maximise and leverage accessible data (structured input-output data pairs and unstructured expert knowledge) to support ENP monitoring, initial screening, and exposure assessment. The developed decision support system could aid to reduce the costs associated with experimental testing and support the establishment of robust frameworks for nano-safety. This is necessary to balance the advancements in nanotechnology and long-term environmental protection. Additionally, the efficiency of developed models can be extended to other ENPs and readily scaled when new and more information becomes accessible without having to reconstruct the frameworks of these models.

Contents

List of figures

List of tables

Nomenclature

Acronyms

ML	Machine learning
FL	Fuzzy logic
AI	Artificial Intelligence
FIS	Fuzzy Inference system
ECs	Emerging contaminants
ENMs	Engineered nanomaterials
ENPs	Engineered nanoparticles
MF	Membership function
MATLAB	Matric laboratory
HDD	Hydrodynamic diameter
IS	Ionic strength
MECs	Measured environmental concentrations
MFA	Material flow analysis
ANN	Artificial neuron network
nm	Nanometre
ANFIS	Adaptive neuro-fuzzy inference system
NOM	Natural organic matter
nZnO	Zinc oxide nanoparticles
CC	Correction coefficient
PECs	Predicted environmental concentrations
PZC	Point of zero charge
SD	Standard deviation
NSE	Nash Sutcliffe efficiency
MAE	Mean absolute error
RMSE	Root mean square error
GT	Gamma test

RFR	Random forest regression
RA	Risk assessment
HM	height method
SVM	Support vector machine
KNN	k-nearest neighbours
GRC	Grey relation coefficient
MOM	mean of maximum
RFPI	Random forest permutation importance
ReLU	Rectified linear unit
GRG	Grey relation grading
WWTPs	wastewater treatment plants
IQR	Interquartile range
KBS	Knowledge base system
COG	Centre of gravity
MNEPs	Multi-enabled nano-products
ARC	Analytic Research Consulting
MEC	Environmental concentration
ENPs	Engineered nanoparticles
OECD	Organisation Economic Cooperation and Development
REACH	Registration, Evaluation, Authorization, and Restriction of Chemicals
MDASR	Meta-data analysis and systematic review
WOE	Weight of evidence
SOE	Strength of evidence
PAIM	permutation accuracy importance measurement
XGBootFI	extreme gradient boosting feature importance
EBP	Evidence-based procedures
SQA	Semi-Quantitative Analysis,
DTC	Decision tree classifiers

Symbols

$\overline{w_i}$	Weighted firing strength
A_i, B_i	Linguistic value: Low, Medium or High
O_j^i	ith Layer
R^2	Coefficient of determinant
R_i	ith Rule
f_i	ith linear output function
n_k	Number of k, k= MF, Rules, Inputs
p_i, q_i	ith consequent parameters
w_i	Firing strength
μ_{A_i}, μ_{B_i}	ith Membership function
x, y	Inputs
σ, c	Premise parameters
ζ	Zeta potential
$\in$	Element off
$\|\ldots\|$	Euclidean distance
$\notin$	Not element
z^*	Deffuzified output
R	Number of the rules
w_j	Ouput value in the j subset
ξ	GRC
arg	Argument
$\hbar_k$	Hidden layer
$x_{norm,i}$	is the ith normalized datapoint
$\hat{y}_I$	ith de-normalized predicted output
ω^2	Regulation term
$\xi_i \xi_i^*$	Slack variables.
α_i *and* α_i^*	Lagrange multiplies
σ	Sigma or standard deviation
Γ	Gamma statistic
γ_i	Grey relation grade

ϵ	Residual terms
β_i	Regression coefficients
t_i	Observed output
μ	Mean
Z and K	Mean of the predicted and observed output

Chapter 1. Introduction

1.1 Background

The use of substances that constitute a peripheral dimension in the geometric range of 1 to 100 nm, has been recognised as a new frontier for innovative research and consequently has transformed modern science over the past decades (Domercq et al., 2018; Lai et al., 2018; Mahaye et al., 2021). Over 9,280 multi-enabled nano-products (MNEPs) in a range of industries, including biomedical, energy, agriculture, medicine, and the food business, are listed in the Nanotechnology Products Database (NPD) published in 2021 (available at https://product.statnano.com, visited in April 2024). Further, the global nanoscience market is expected to expand by $8.5 billion between 2021 and 2026, according to Industry Analytic Research Consulting (ARC) (available at https://www.indusgtryarc.com /Report/16067/nano chemicals, visited in April 2024).

Apart from the rapid expansion, and increased demand for MNEPs across a range of engineering platforms: it is well-recognised that the fast development of nanotechnology will subsequently increase the environmental concentration of engineered nanoparticles (ENPs) (Abbas et al., 2020; Leareng et al., 2020). This is evident by various predicted and analytically measured concentrations covered extensively in research by Zhao et al. (2021). For example, the analytically measured environmental concentration (MEC) of nTiO_2 was found in the range of 0.2 to 4 µg/ℓ in Taihu Lake in China, the Meuse River in the Netherlands, and the Salt River in the United States (Peters et al., 2018; Venkatesan et al., 2018; Xiao et al., 2019). Additionally, the modelled concentration of nZnO in surface waters was estimated to be between 0.01 and 0.1 µg/ℓ (Gottschalk et al., 2013) and to surpass 0.150 µg/ℓ in Europe (Dumont et al., 2015).

Given the risks of ENPs are generally unknown, their continuous emission, and ubiquitous occurrence in aquatic systems, continue to be a subject of scientific interest and scrutiny (Hou et al., 2018; Musee et al., 2014). The novel properties and unique behaviour of ENPs which offer a multitude of functionalities across a range of fields have been identified as the primary causes of biochemical, morphological, structural,

and/or physiological deleterious effects induced on microorganisms (Hou et al., 2018; Mahaye et al., 2017; Thwala et al., 2016). Studying the fate, and behaviour of ENPs, is therefore essential to properly address the current environmental safety concerns as well as to foster the sustainability of nanotechnology.

In the past 1-2 decades several approaches including empirical, and mechanistic modelling have been reported (Klein et al., 2016; Meesters et al., 2019; Nyangiwe and Ouma, 2019). The status quo, challenges and perspective of these techniques are well discussed in several studies (Domercq et al., 2018; Williams et al., 2019). Probalistic models or basic material flow analysis (MFA) are quite useful for examining the flow and concentration in various compartments (Bundschuh et al., 2018). Mechanistic models are valuable for the mathematical description of behaviour phenomena. However, MFA modelling studies have challenges concomitant to large data uncertainties and model input parameterisation (Domercq et al., 2018; Musee, 2018; Nowack, 2017). The mathematical formula used in MFA was originally designed for bulk materials (Arvidsson et al., 2011; Dale et al., 2015; Nowack, 2017). No recognition and consideration of the unique transformations of species (particulates, colloids, and ions) and the interfacial reactions that occur on ENPs (Zhang et al., 2019). On the other hand, the mechanistics are described as highly complex and technical, and often applicable to specific physicochemical properties such as spherical nanoparticles (Hristozov et al., 2016).

In addition to the existing modelling studies, experimental data has been generated to elucidate the transformation behaviour of ENPs; yet, the data is regarded as knowledge- and information-poor. This is because the current data has characteristics that include heterogeneity, data gaps, highly multifaceted relationships (not easily discernible data variables), fragmentation, the multiplicity of data, and contradictory information (Ban et al., 2020; Furxhi et al., 2019a; Mirzaei et al., 2021; Peng et al., 2020). These attributes are concomitant with the paucity of consistent reporting protocols, high variability of instruments used for measuring various parameters, and deficiency of appropriate experimental controls, among others (Ban et al., 2018; Basei et al., 2019a). As a result, even though the volume of data is increasing, the ability to derive intelligent decisions has diminished as illustrated using the pyramid scheme in Figure 1.1. Therefore, there is an urgent need for the investigation of advanced cutting-

edge tools to decipher and provide long-term solutions that seek to address the exposure and risk assessment of ENPs.

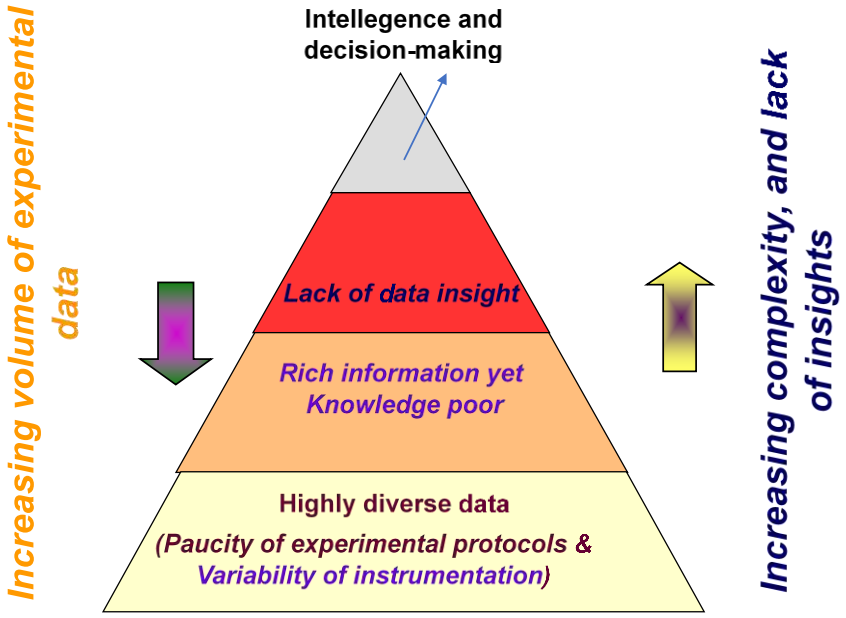


Figure 1. 1. Pyramid scheme representing the existing knowledge (Ban et al., 2018).

1.2 Rationale for Research

In recent years, there have been growing concerns about the potential impact of ENPs on the environment, as well as recognition of the need to understand their fate and likely environmental implications as evidenced by increasing experimental data (Abbas et al., 2020; Venkatesan et al., 2018). This is due to the rising innovative use of nano-enabled products and the need to balance between simulating economic growth and societal demands for safe chemical products as per the guidelines on Safety testing and assessment of Manufactured Nanomaterials by the Organisation for Economic Cooperation and Development (OECD) (OECD, 2016). However, in the past decades, even though the fate and behaviour of ENPs in the aquatic system have been studied, there is still a knowledge gap or research niche on the role of modelling specifically the applications of data modelling to assist with decision-making.

Considering that it is challenging and impossible to cover and cope with higher costs concomitant experimental testing of fate and behaviour using different aquatic permutations for every ENP introduced on the market (Concu et al., 2017; Takahashi

and Takahashi, 2019). There is a need to investigate alternatively non-testing methods that can expedite decision-making and minimize experimental testing as recommended by European Chemical Agency regulations on the Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) (El Mahdi and Aziz, 2018). Therefore, with advancements in big data and/or artificial intelligence (AI), this work describes the use of machine learning (ML) and knowledge base systems (KBS) to discover scientific information and provide a coherent understanding of the mechanisms that drive the transformation of ENPs as well as the development of intelligent decision system.

ML is a sub-field of AI that employs the learning-by-example approach (Alpaydin, 2020; Mohri et al., 2018). ML trains computers to extract multidimensional, hidden information in a domain without the predefined relationship between the input and output parameters (Blum and Langley, 1997; Jordan and Mitchell, 2015). ML include artificial neural networks (ANN) (Pham et al., 2019; Rosenblatt, 1958), support vector regression (SVR) (Chen et al., 2019; Cortes and Vapnik, 1995), adaptive neuro-fuzzy systems (ANFIS) (Choubin et al., 2018; Jang, 1993), and random forest (RF) techniques (Breiman, 2001; Meng et al., 2018; Stafoggia et al., 2019). Recently, ML has shown a growing interest across various scientific fields and engineering systems, as evidenced by the increasing number of research studies reported on environmental domains (Ban et al., 2018; Furxhi et al., 2019a; Mirzaei et al., 2021). Many factors, 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, have contributed to the growing use of ML (Glaubitz et al., 2022; Jordan and Mitchell, 2015; Sun and Scanlon, 2019).

Additionally, KBS is another form of AI that is generally beneficial in situations where quantitative scientific information or data is lacking, as the case in nanoecotoxicology (Amirshenava and Osanloo, 2019; Grella et al., 2019; Obiedat and Samarasinghe, 2016). Among many factors, the ability of KBS like Mamdani-fuzzy logic to combine heuristics and expert knowledge helps to solve complex problems (Aqlan, 2016; Pepa et al., 2020). Therefore, the use of advanced modelling methods applied in this study is envisaged to provide a foundation to leverage the strengths of data modelling to

assist in decision-making within nanoecotoxicology domain and can be extended to numerous other environmental problems.

1.3 Research Aim and Objectives

1.3.1 Aims

The aim of this study was the application of data modelling techniques in development of computer-based intelligent systems to support decision-making in dealing with ENPs. At the same time, the study aims to provide a coherent understanding of the mechanisms and interactions that underpin the fate and behaviour of ENPs in aqueous environments as summarised using the schematic diagram in Figure 1. 2.

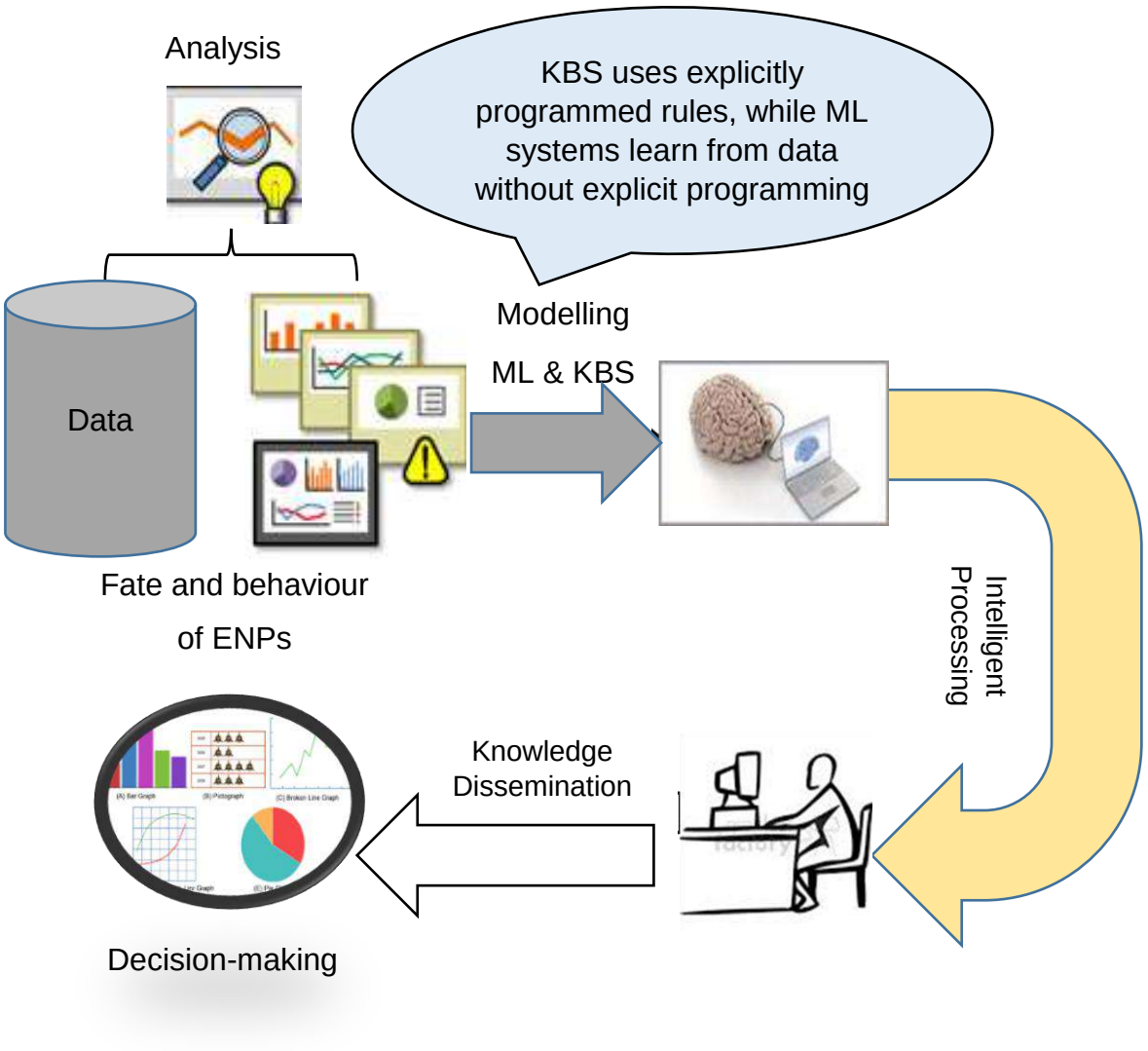


Figure 1. 2. Schematic diagram showing the research study aim (Glaubitz et al., 2022).

1.3.2 Objectives

Herein to achieve the research study aims, these objectives were followed;

1. Critically examine the literature on the fate and behaviour of ENPs in aqueous environments using the meta-data analysis and systematic review (MDASR) process
2. Supervise various linear and non-linear ML algorithms using Python and R program language from historical data
   2.1 Extract deeply multidimensional information and learn underlying patterns
   2.2 Rank and identify the preponderant variables that influence the behaviour of ENPs in aquatic systems
   2.4 Train ML to predict the transformation of ENPs in freshwater-like systems using experimentally reported physicochemical and water chemistry input variables
   2.5 Evaluate, and compare the robustness and the prediction accuracy of ML using several statistical metrics
3 Developing Knowledge-Based Systems using MATLAB and Excel
   3.1 Apply evidence-based procedures and Occam's Razor parsimonious concepts to select variables based on the weight of evidence (WOE) and strength of evidence (SOE)
   3.2 Develop a parsimonious hierarchical framework that maps the interrelationship among multivariate inputs to selected exposure output indicators.
   3.3 Encode the domain knowledge using scores and rules to develop intelligent KBS that support decision-making at the expert level
   3.4 Evaluated the developed KBS using case studies

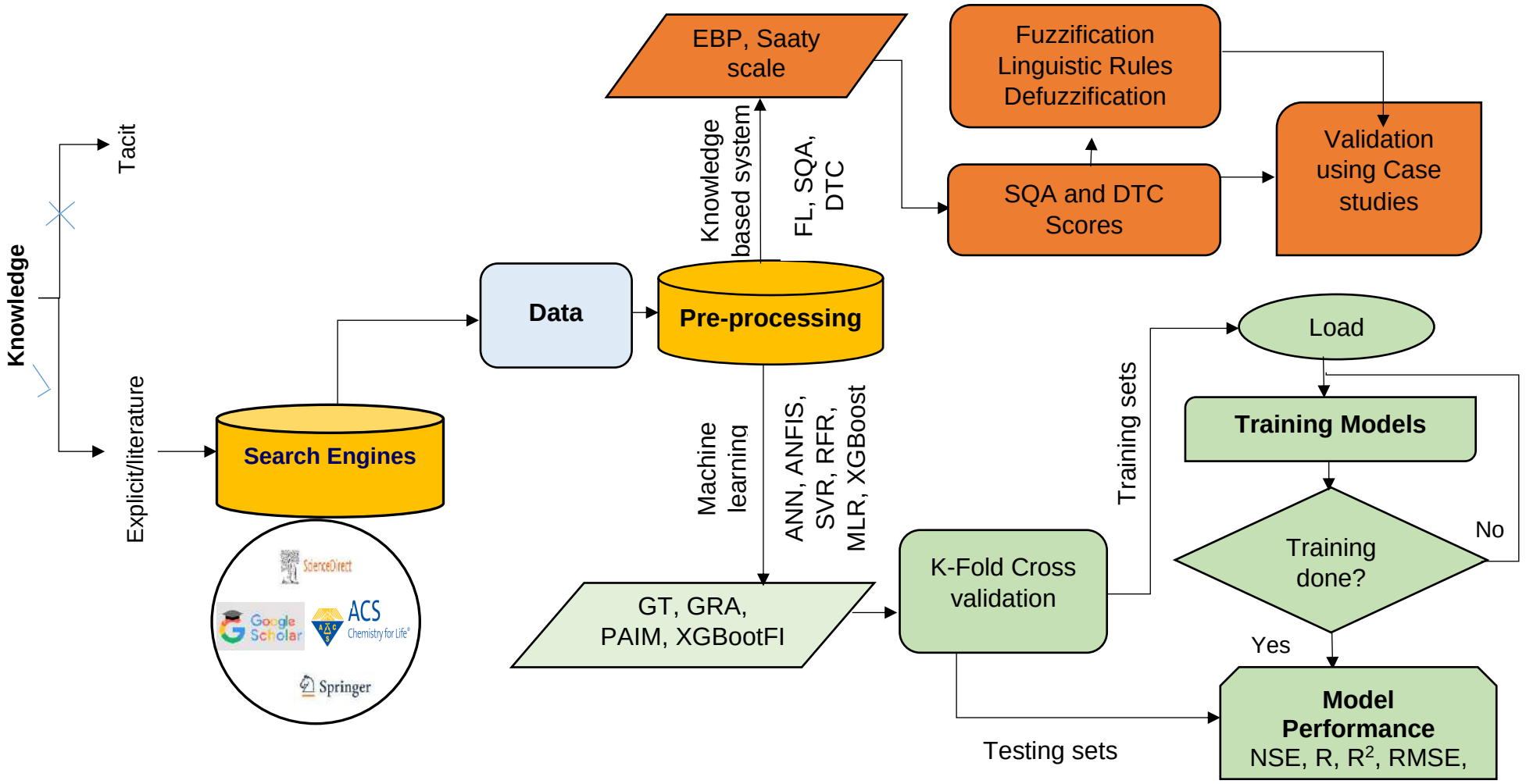


Figure 1. 3. Schematic flow representing the study design

1.4 Study design

Figure 1. 3 shows the schematic flow representing the study design. Presently, the lack of uniform and structured data in publicly available reader-friendly databases that can be easily analysed is a defining feature of the nanoecotoxicology field (Ban et al., 2018; Goldberg et al., 2015). Knowledge in the domain can be represented as tacit and/or explicit in literature. In this study, secondary data was solicited from literature research using the MDASR procedure (Foley et al., 2018; Gurevitch, 1993; Haidich, 2010). In this process, both quantitative and non-quantitative data were collected from the literature and rigorously assessed to create a database and expert knowledge. Subsequently, various ML approaches (ANN, SVR, RFR, ANFIS, MLR, XGB, and KNN) were supervised to develop predictive algorithms. On the other end, knowledge was encoded using rules and scores to develop fuzzy logic- and semi-quantitative based computer models, respectively. KBS uses explicitly programmed rules, while ML systems learn from data without explicit programming. The functionality and applicability of these approaches have been demonstrated using nZnO, nAg and $nTiO_2$ as case studies since it was impossible to cover all the ENPs on the market.

1.5 Thesis Structure

The thesis is divided into the following chapters.

Chapter 1: This chapter summarises the background of ENPs and provides the rationale for the study

Chapter 2: This chapter covered salient factors influencing ENP emission including manufacturing and applications. Additionally, a comprehensive review of environmental fate models including mechanistic, and data models is provided

Chapter 3: This chapter provides details on the several methods and materials used for developing the ML predictive algorithms and KBS.

Chapter 4: This chapter provides details on the application of ML approaches to estimate the aggregation of zinc oxide and titanium dioxide nanoparticles in freshwater-like systems.

Chapter 5: This chapter provides details on developing predictive models using ML approaches to estimate the dissolution of zinc oxide in aqueous environmental systems.

Chapter 6: This chapter provides details on developing the model for screening the fate and behaviour of the ENPs in aquatic systems using semi-quantitative analysis and decision tree classifiers

Chapter 7: This chapter provides details on developing the fuzzy decision-making system to evaluate the fate and behaviour of metal-based ENPs in aquatic systems

Chapter 2. Literature review

2.1 Introduction

New technologies such as nanotechnology are being introduced into the market, which helps to improve material performance and boost economic growth (Domercq et al., 2018; Williams et al., 2019). However, in turn, this has led to the emission of tons of unwanted chemicals that seriously endanger natural lifeforms (Alharbi et al., 2018; Liu et al., 2017). ENPs have been one of the major chemical pollutants of environmental concern in the last two decades (Domercq et al., 2018) including personal care products (PCPs) (Liu et al., 2020), pharmaceuticals (Rivera-Utrilla et al., 2013), disinfectants (e.g., triclosan and triclocarban) (Musee, 2018), sanitizers (Musee et al., 2020), microplastics (MPs) (Du et al., 2020; Sun et al., 2021), etc.

ENPs, in particular, have garnered great scientific interest and are materials of interest in this investigation. Their global market is estimated to surpass USD 125 billion by 2024 (https://www.bccresearch.com/marketresearch/nanotechnology, visited in April 2024). However, apart from the rapid expansion, and increasing demand for multi-enabled nanoproducts across a range of engineering platforms; following their widespread applications, approximately 7% of ENPs sink into aquatic systems. Once they enter the aquatic system they pose a serious environmental concern to aquatic biota and food web (Bundschuh et al., 2018; Mahaye et al., 2021; Mahaye and Musee, 2023). For example, the estimated release of nAg and nZnO into surface water is approximately 4.9–1700 tons per year (Rajkovic et al., 2020). Furthermore, according to Giese et al. (2018), environmental concentrations (ECs) in freshwater are expected to rise exponentially to hundreds of ng/ℓ in 2050. Therefore, it is important to address their risks and environmental health and safety (EHS) concerns for the sustainability of nanotechnology (Zhao et al., 2021).

Risks of environment contaminants can be defined by two important components namely; exposure and hazardous components (dose-response relationships). Exposure assessment (EA) measures the extent or probability of a contaminant's exposure potency to organisms that are not the intended target, such as aquatic vertebrates and invertebrates (Tolaymat et al., 2015; Topuz and van Gestel, 2016). This process requires an understanding of ENP production, distribution, and mechanisms that underpin their bioavailability and bioaccumulation (Meesters et al.,

2013; Musee, 2018). ENP production and distribution have been thoroughly addressed elsewhere (Domercq et al., 2018; Williams et al., 2019). However, there is a knowledge gap in understanding their fate and behaviour, despite the increasing volumes of experimental data. Thus, in this chapter, we begin by briefly discussing the progressive application, possible emission pathways, environmental concentrations as well as the experimental data on the fate and behaviour of ENPs in aquatic systems. Then lastly, review the environmental modelling techniques with a specific interest in mechanistic and data modelling.

2.2 Engineered nanoparticles

ENPs are man-made material(s) that contain one or more closely bound substances whose geometrical dimensions are at least partially within the 1-100 nm size range, and whose components are molecules and atoms (Buzea et al., 2007). ENPs possess distinct and adjustable physicochemical characteristics, such as a high surface area-to-volume ratio (per unit volume or mass) and quantum effects, which set them apart from their bulk counterparts (Hochella et al., 2019; Sengul and Asmatulu, 2020). These materials are synthesised by either top-down and/or bottom-up techniques, such as liquid-phase synthesis, chemical and physical vapour deposition, etc., (Ealia and Saravanakumar, 2017; Hong et al., 2006; Swihart, 2003).

ENPs are classified into several categories which include the inorganic-based nanoparticles (INPs) ($nTiO_2$, nAg, nZnO, nAl_3O_2, nFe_3O_2, and $nCeO_2$), carbon-based (carbon nanotubes (CNT), fullerenes (C60) and their derivatives, carbon black, graphite nanoparticles, graphene nanoparticles and graphene oxide), and quantum dots, etc. (Klaine et al., 2008; Yokel and MacPhail, 2011). Carbon nanotubes (single- and multi-walled carbon nanotubes) are used in tennis rackets, batteries, and the water treatment process (Adam and Nowack, 2017). Quantum dots find application in semiconductors, films, fluorescent dyes for photography, bio-probes and biosensors, and imaging agents (Algar et al., 2010; Frigerio et al., 2012).

INPs find applications in various fields as described in Table 2.1, and account for about 82–87% and 89–97% of ENPs released into soil and water, respectively (Keller et al., 2013). Strong coordinate bonds, a high refractive index, resistance to discoloration, enhanced photo-catalysis, high mechanical strength, chemical inertness, surface super-hydrophobicity, and antibacterial properties are just a few of the many diverse

physicochemical characteristics that INPs possess (Besinis et al., 2014; Guo et al., 2019; Vance et al., 2015). About 304,000 metric tonnes of silicon dioxide ($nSiO_2$) titanium dioxide ($nTiO_2$), iron oxides (nFeOx), aluminium oxides (nAlOx), zinc oxide (nZnO), and cerium dioxide ($nCeO_2$) nanoparticles were estimated to be produced annually (Medina-Velo et al., 2017).

Among INPs, $nTiO_2$ has the highest production volume of 100-1000 tons per year (60,000 tonnes per year), followed by nZnO (nZnO; 10,000 years) and nAg (Bossa et al., 2017; Keller et al., 2013; Piccinno et al., 2012; Sun et al., 2014). $nTiO_2$ production volumes exceed 27.397 tons per day, with production volumes ranging from 0.27397 to 2.7397 tons per day for nZnO (Piccinno et al., 2012). The categories of cosmetics and paint & coatings account for 59% and 13% of $nTiO_2$, respectively (Grande and Tucci, 2016). This is because of its physicochemical properties, including its greater refractive index and its ability to absorb ultraviolet (UV) light. Moreover, nZnO has received significant attention in various fields ranging from automotive, biomedical, energy, and electronic products (Foss Hansen et al., 2016; Grillo et al., 2018; Sengul and Asmatulu, 2020). The unique characteristics of nZnO, including a wide band gap (3.37 eV), a large exciton binding energy (60 m eV at 300k) (Debanath and Karmakar, 2013), antimicrobial activity (Sirelkhatim et al., 2015), piezoelectric and pyroelectric properties (Parihar et al., 2018) are contributing attributes. The antibacterial qualities of nAg can be attributed to the expansion of its application in consumer products, including clothes, on the global market (Jahan et al., 2024).

Table 2. 1. Various applications of metallic nanoparticles in different fields

Metallic ENPs	**Application**	**References**
Ag	Consumer products include electronics, cosmetics, household appliances, textiles, and food production as well as biomedical applications such as antimicrobial agents, drug delivery, molecular imaging, biomedical sensing, and even cancer photodynamic therapy	[1-5]

Al_2O_3	Fuel cells, polymers, paints, coatings, textiles, biomaterials, batteries, adsorbent, grinding, catalysis, polishing abrasives	[6-9]
TiO_2	Paints, plastics, cosmetics, personal care products, food additives and drug delivery agents, coatings, papers, inks, medicines, pharmaceuticals, food products, toothpaste	[10-20]
ZnO	Catalysis, paints, wave filters, UV detectors, transparent conductive films, gas sensors, solar cells, sunscreens, and cosmetic products	[21-25]
IO	Drug delivery, magnetic resonance imaging, thermal ablation therapy, in vivo cell tracking, magnetic separation of cells or molecules, and remediation of different environmental contaminants such as heavy metals, chlorinated organic solvents	[26-28]

IO: iron oxide, TiO_2: titanium dioxide, ZnO: zinc oxide, Ag: silver; Copper oxide; Au; Gold; aluminum oxide. (Adopted from Kurma et al., 2018). [1]: Akter et al. (2018), [2]: Cameron et al. (2018), [3]: Chen et al., (2015), [4]: Foldbjerg et al. (2011), [5]: Guo et al. (2019), [6]: Future Markets (2013), [7]: Kim et al. (2010) and [9]; Poborilova et al. (2013), [10]: Besinis et al. (2014), [11]: Neal et al. (2011), [12]: Nthwane et al. (2019), [14]: Qi et al. (2013), [15]: Ranjan and Ramalingam, (2016), [16]: Shandilya et al. (2015), [17] Shi et al. (2013), [18]: Weir et al. (2012), [19]: Windler et al. (2012), [20]; Yin et al. (2012), [21]: Choi et al. (2018b), [22]: Coll et al. (2016), [23]: Ghasemi and Rohani (2019), [24]: Guan et al. (2012), [25]: Huang et al. (2010), [26]: Feng et al. (2018), [27]: Guerra et al. (2018), [28]: Naqvi et al. (2010).

2.3 Source and Emission of ENPs into the ecology

Recent developments in the production of ENPs have resulted in ubiquitous occurrence in ecology systems. ENPs exposure in ecological systems is attributed to two main sources namely: natural processes, which include the chemical deterioration of naturally occurring minerals such as continental crust rocks and/or man-made activities described in Figure 2.1 (Johnson et al., 2011). Nevertheless, there is a paucity of data regarding the release of natural processes. On the other hand, three scenarios are generally taken into consideration to examine the emission of ENP into aquatic systems, namely: (i) release during the production of raw materials and nano-enabled products; (ii) release during use; and (iii) release following the disposal of products containing NP (waste handling). During use, the emission of ENP into aquatic

systems can be either directly via effluent in underperforming WWTPs or indirectly through leaching in agriculture or sludge disposed in landfills (Bundschuh et al., 2018).

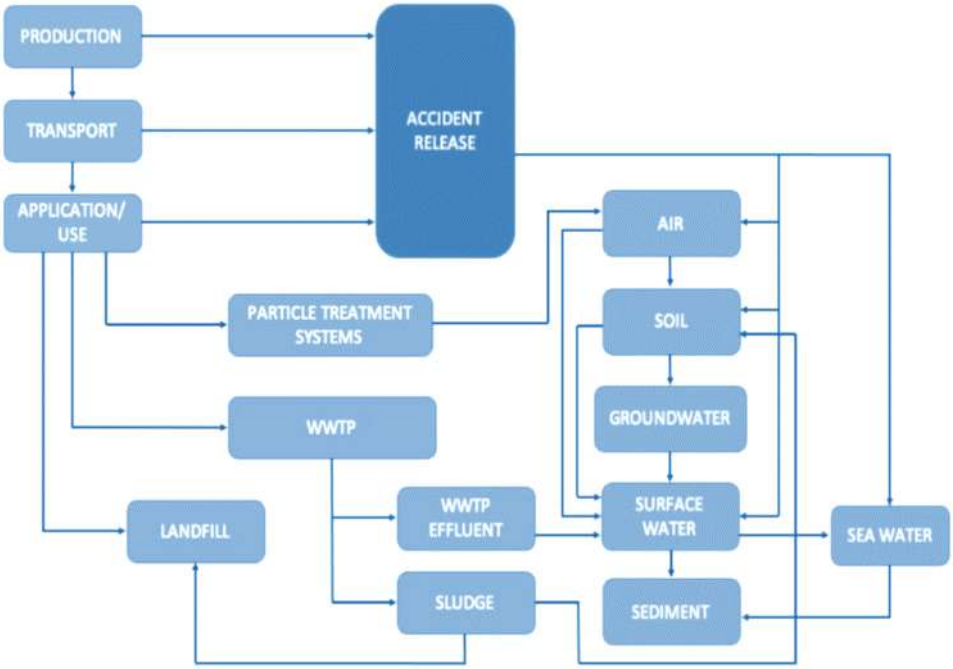


Figure 2. 1. Emission of ENPs into the aquatic systems. Adapted with permission from (Ramirez et al., 2022). Copyright © 2021, by the authors. Used under a Creative Commons (CC BY) License.

In various compartments including aquatic systems the ECs of ENPs have been determined using analytical measurements and predictive modelling (Zhao et al., 2021). Probalistic models or basic MFA are used in numerous studies to examine the flow and concentration in various compartments. On the other hand, analytical measurements use spot-grabbed aquatic samples or simulated exposure medium to trace the ECs of ENPs from variant consumer products (Al-Kattan et al., 2015; Bossa et al., 2017; Shandilya et al., 2015). The characterisation and quantification of ENPs in natural environments use several instrumentations such as transmission electron microscopy/energy dispersive X-ray analysis (TEM/EDX), inductively coupled plasma-mass spectrometry (ICP-MS) (Keagi et al., 2008), electro-thermal atomic absorption spectrometry (ET-AAS) (Li et al., 2016), asymmetric-flow field flow fractionation with inductively coupled plasma–mass spectrometry (AF4-ICP-MS) (Hoque et al., 2012), and especially, single particle inductively coupled plasma-mass spectrometry (sp-ICP-MS) (Peters et al., 2018; Wu et al., 2020; Xiao et al., 2019).

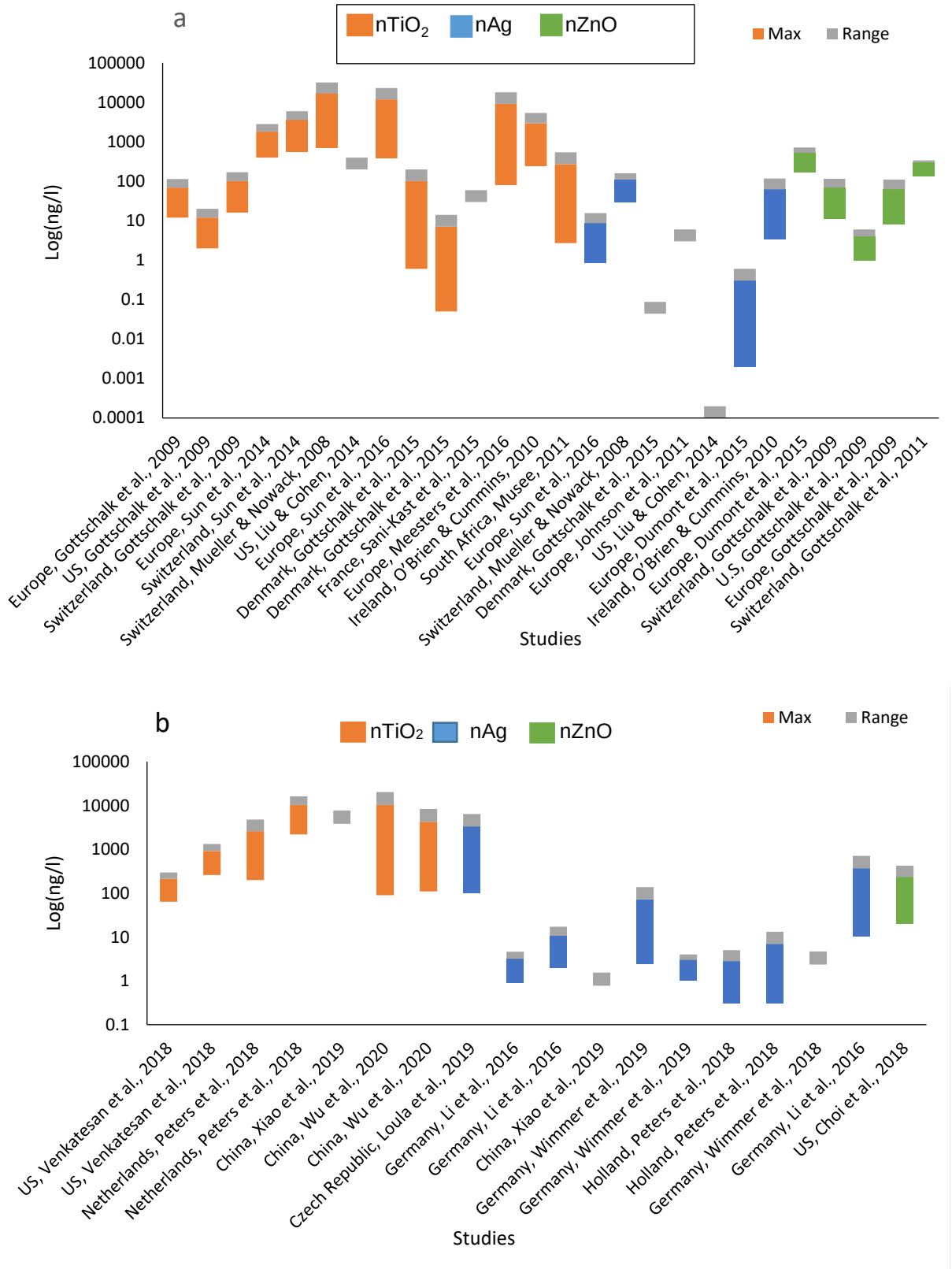


Figure 2. 2. Environmental concentrations in natural systems (a) predicted and (b) measured (Zhao et al., 2021).

Both predicted environmental concentrations (PECs) and measured environmental concentrations (MECs) in natural freshwater systems are summarised in Figure 2.2 using a logarithm scale. In Figure 2.2(a) the PEC of $nTiO_2$ in surface water was 0-30 ng/ℓ in the Rhône River, France (Sani-Kast et al., 2015). The nAg in surface water in Europe was 0.87-7.84 ng/ℓ (Sun et al., 2016) and 0.03-2.79 ng/ℓ in the US (Giese et al., 2018). Additionally, about 0.01 to 0.150 µg/ℓ of nZnO concentration was estimated in surface waters (Dumont et al., 2015; Gottschalk et al., 2013). In Figure 2.2(b) the MEC corresponds to the PECs. In the Meuse and the Ijssel rivers (Peters et al., 2018), in Taihu Lake (Wu et al., 2020), a Golden Colorado stream (Reed et al., 2017), river water (Neal et al., 2011) the concentration of $nTiO_2$ was measured in the range of 0.09 to 10.2 µg/ℓ (Xiao et al., 2019). The nAg was found to be between 2.0 - 8.6 ng/ℓ in the Meuse, Ijssel River in the Netherlands and Taihu Lake in China (Peters et al., 2018; Xiao et al., 2019).

2.4 Surface transformations of ENPs in the aqueous environments

ENPs undergo a variety of transformation processes in natural systems (Abbas et al., 2020). These transformation processes include organic matter adsorption (Danielsson et al., 2018), aggregation (Ottofuelling et al., 2011; Wagner et al., 2014), and dissolution (Bian et al., 2011; Odzak et al., 2017), among others. The irreversible gathering of colloid particles into sizable agglomerates based on the effectiveness of their collisions is known as aggregation (Dwivedi et al., 2015; Grillo et al., 2015). Brownian motion (perikinetic aggregation) and fluid motion (orthokinetic aggregation) have an impact on the collision frequencies between particles (Hartmann et al., 2014). In natural or complex systems, heteroaggregation—the interactions of ENPs with inorganic ions such as Ca^{2+}, and natural colloids such as NOM—predominates over homo-aggregation (Praetorius et al., 2014). According to Stoke's theorem, heavy agglomerates settle more quickly because of the higher gravitational field (Hartmann et al., 2014).

Moreover, the release of water-soluble ionic species from ENP into an aqueous solution is known as dissolution. This results in a decreased concentration of stable colloidal particles in suspension (Hedberg et al., 2019). The Noyes-Whitney states that because the rate of dissolution is proportional to the surface area, small nanoparticles dissolve more quickly (Bian et al., 2011; Tangaa et al., 2016). Additionally to this, steric or electrostatic repulsion forces stabilise ENP (Ottofuelling et al., 2011; Wagner et al.,

2014). Interactions with macromolecules, such as organic matter (via adsorption — the adherence of adsorbate molecules to the adsorbent surface, such as NOM) or surface coatings, produce steric repulsion forces. These substances block potential reactions with other substrates or NP-NP interactions by imparting negative charges, thus, producing extremely stable ENPs (Louie et al., 2016; Philippe and Schaumann, 2014).

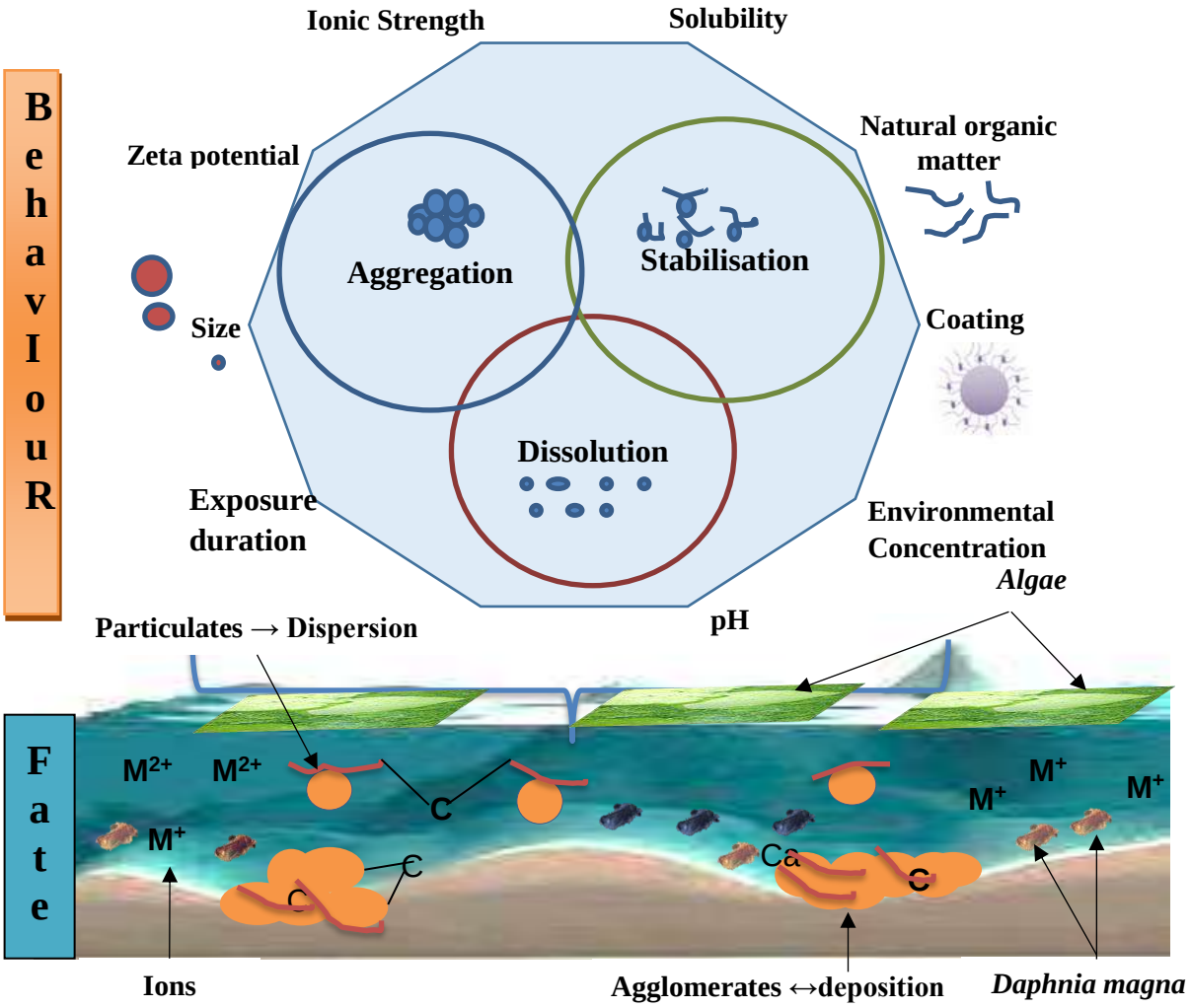


Figure 2. 3. The factors that influence the degree of exposure of ENPs to microorganisms in aquatic systems (Abbas et al., 2020).

These ENP behaviour processes in aqueous environments influence their exposure (e.g., deposition, bioavailability, persistence, bioaccumulation, dispersion, etc.) and, in turn, their toxicological effects as depicted in Figure 2.3 (Jiang et al., 2015a; Wang et al., 2018). High gravitational settling (high deposition rate) results from high aggregation/rapid agglomeration (high collision efficiency). This, in turn, reduces the suspended ENPs in the suspension and toxicity, while the concentration increases in sediments (Chekli et al., 2015; Schaumann et al., 2015). For example, the agglomeration and the deposition processes of $nTiO_2$ reduced the deleterious effects on microalgae *Isochrysis galbana cells* (Hu et al.. 2018). On the other hand, a higher degree of colloidal stabilisation increases ENPs dispersion, bioaccumulation, and persistence of stabilised particles in suspension (Abbas et al., 2020).

Aquatic chemistry (e.g., natural organic matter (NOM), ionic strength (IS), pH, etc.) and their intrinsic physicochemical features influence the ENPs transformation processes in environmental systems (Amde et al., 2017; Philippe and Schaumann, 2014; Wagner et al., 2014). Large amounts of experimental data are being produced, but current terms of reference for investigating the colloidal stability of ENPs are based on vague. Among other things, this is due to the non-standardisation of experimental protocols, variations in exposure media and intrinsic properties, and measurement methods applied in different experiments (Nowack et al., 2015; Selck et al., 2016). As a consequent, this has resulted in numerous contradictions and a lack of understanding regarding the significance of features that influence the transformation processes in aquatic environments for decision-making.

2.5 Modelling

In most environmental domains including nanoecotoxicology the data generated experimentally on the fate and behaviour of contaminants in natural systems are generally recognised as complex and open-ended, characterised by various forms of uncertainty such as poorly structured information, lack of homogeneity, and data gaps (Dagnino et al., 2013; Pang and Coghill, 2015a). The use of modelling to elucidate and study complex systems using an algorithmic or mechanistic approach is considered an alternative conceptual tool capable of supporting decision-making (Cowan, 1995; Di Guardo et al., 2018). Modelling is widely used in a diverse range of fields spanning physics, engineering, chemistry and biology, among others

2.5.1 Mechanistic based modelling

Mechanistic models are based on fundamental laws of the natural sciences, including physical and biochemical principles (Di Guardo et al., 2018). They have a long history of application and are widely used to investigate the behaviour of organic chemical pollutants in many complicated problems (Di Guardo et al., 2018). Fugacity-based models (Webster et al., 1998), SimpleBox (Klasmeier et al., 2006), Multimedia Activity Model for Ionics (MAMI) (Franco and Trapp, 2010), and Sino Evaluative Simplebox-MAMI (SESAMe) models are the most successful prototypes of environmental fate models (EFMs) for organic environmental contaminants (Zhu et al., 2014).

Components of conventional organic-based models have been applied in the field of nanoecotoxicology (Meesters et al., 2014a). These include multi-media fate models, which integrate material flow analysis and colloidal chemistry theory and kinetics based on the principles of attachment efficiency, collusion frequency, fractural dimension, Smoluchowski-Stokes mechanistic, and first-rate equations. Examples include NanoDUFLOW (Klein et al., 2016) and SimpleBox4nano (SB4N) (Meesters et al., 2014a), GWAVA (global water availability assessment) (Dumont et al., 2015), and FINE (Money et al., 2012).

The SimpleBox4Nano was applied to estimate concentration in surface water, soil, and sediments (Jacobs et al., 2016). Money et al. (2014) showed that the prediction accuracy of the FINE was approximately 70% compared to the traditional baseline model using the nAg. More successful applications of dynamic models such as NanoDUFLOW and SB4N have been described by (Klein et al., 2016) and (Meesters et al., 2016, 2019), respectively. However, despite the benefits associated with these mechanistic approaches—such as the reduced reliance on experimental data for assessment and calibration, and making good predictions outside the range of previously used input values — the drawbacks include involved intricate mathematical functions that are utilized to explain fate (Hristozov et al., 2016).

2.5.2 Data based Modelling

Data models primarily involve the manipulation of data or knowledge using a myriad of scientific methods, processes, and algorithms to formulate decision support systems (DSS) or intelligent decision support systems (IDSS) (García-Diéguez et al., 2015; Yazdani et al., 2017). These are designed to enable computers to

perform tasks that would require human intelligence. The use of data modelling has gained prominence across various fields together with the development of digital computers, and advancements in AI research. In this study, we discuss two types of recognised data models namely i.e., structural and fully orientated data-driven models such as ML, and rule base models such as KBS. The nature of data attributes and representation in a given problem is salient to the choice of the model/ algorithms (Camastra et al., 2015; Yazdani et al., 2017). Figure 2.4 describes the characteristics of ML and KBS based on representation, learning, adaptability and handling uncertainty.

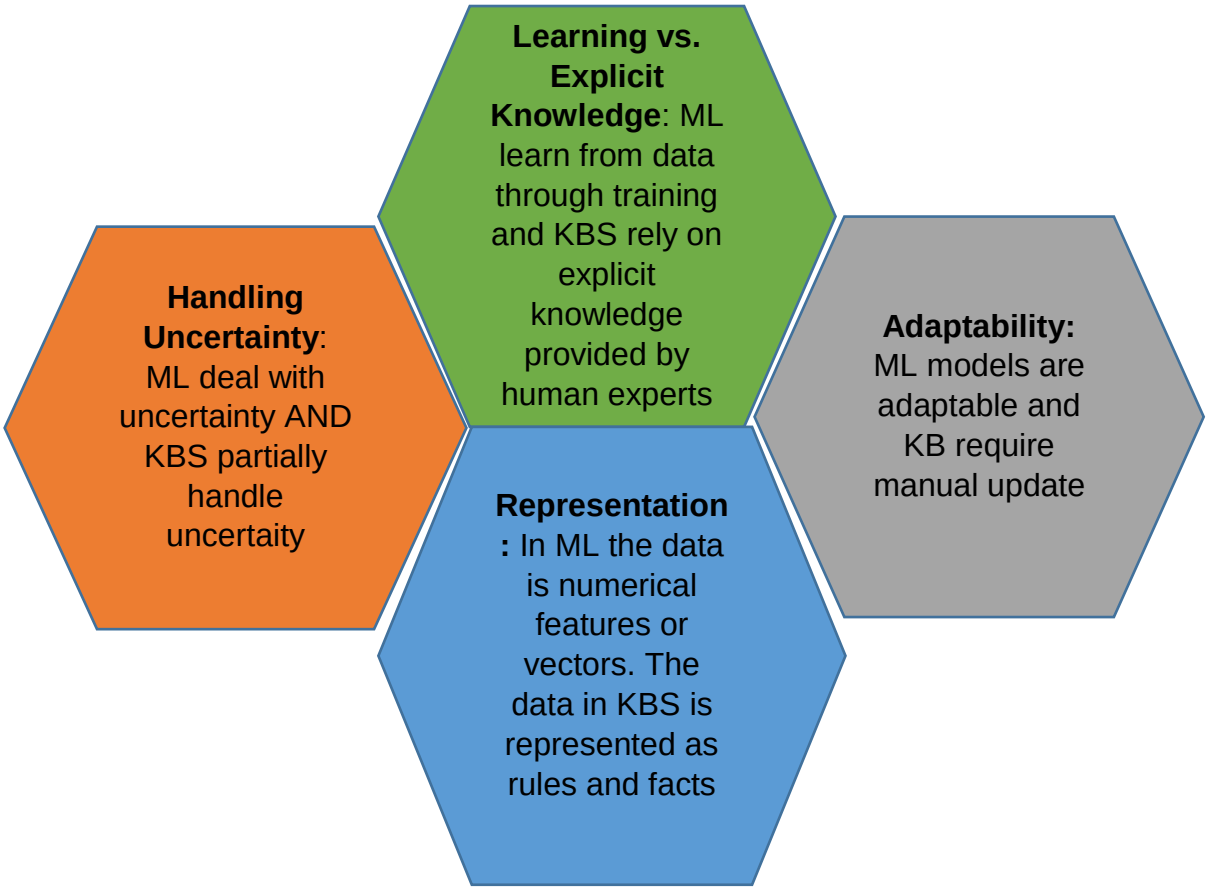


Figure 2. 4. Fundamental difference between KBS and ML (Rosati et al., 2023).

2.5.2.1 Machine learning

ML is a sub-field of AI and a multidisciplinary domain that uses a learning-by-example paradigm (Alpaydin, 2020; Mohri et al., 2018). ML trains computers to extract hidden, deeply multidimensional information where no fully satisfactory algorithm is available (Blum and Langley, 1997; Jordan and Mitchell, 2015). The ML can be broadly categorised into three groups; supervised, non-supervised and reinforcement learning

as described in Figure 2.5. Reinforcement learning enhances the model performance by interacting with the environment and is widely used in games (Peng et al., 2021).

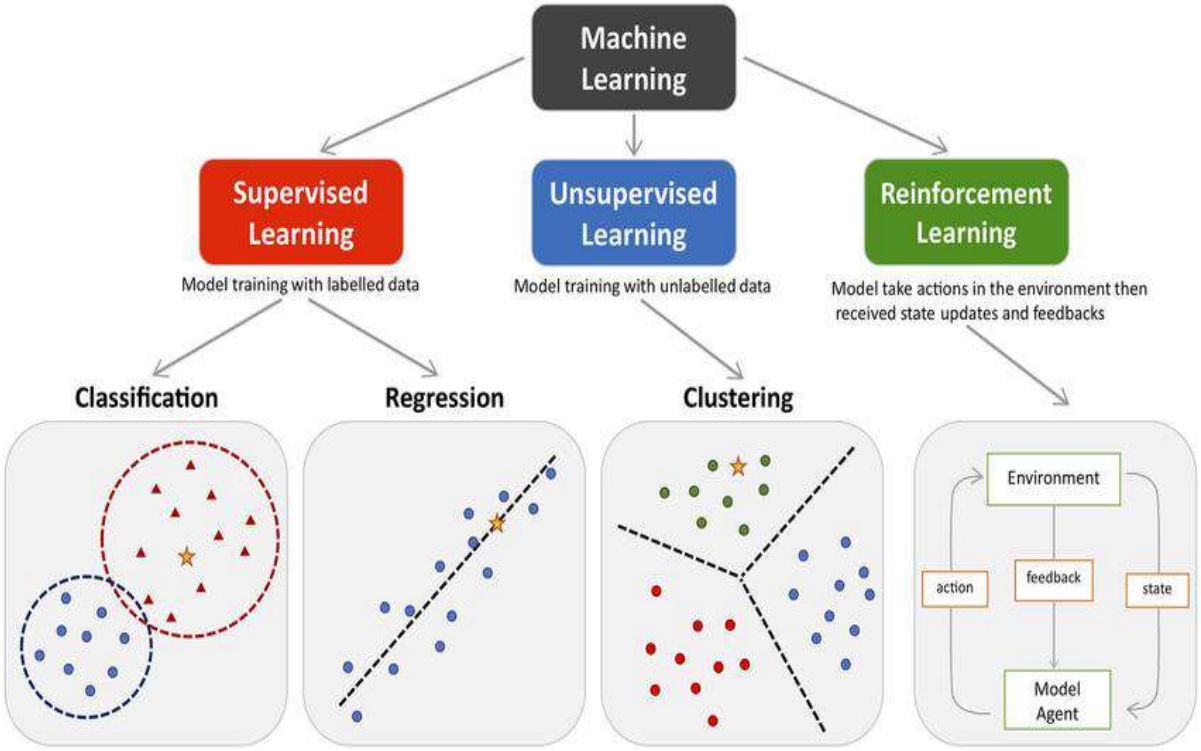


Figure 2. 5. Various categories of ML. Adapted with permission from (Peng et al., 2021). Copyright © 2021 by the authors. Used under a Creative Commons (CC BY) License.

Recently, ML tools have witnessed a growing interest across various fields including nanoecotoxicology to support the decision-making process. The rise in the use of ML can be attributed to manifold reasons, such as their effectiveness in anomaly detection, handling data defined by uncertainties, ambiguities, and non-linearity as well as their high learning ability, good error tolerance, less computer code, and easily updated (Furxhi et al., 2020; Miller et al., 2018; Sun and Scanlon, 2019). Miller et al. (2018) summarise applications of ML in toxicology. In a review paper by Furxhi et al. (2020), the advantages of ML in nanotoxicology are discussed. Further, the environmental benefit and future direction prospects are described by Jordan and Mitchell, (2015) and Sun and Scanlon, (2019).

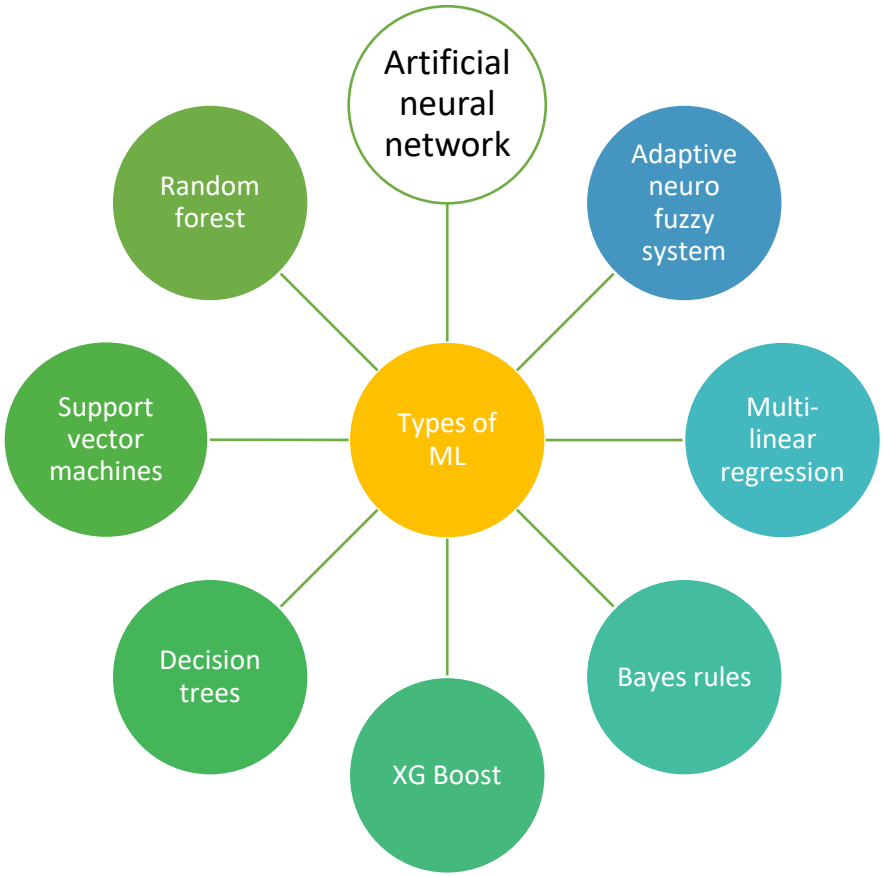


Figure 2. 6. Widely applied ML techniques (Dong et al., 2022).

ML tools in Figure 2.6 are categorised into various classes such as trees, e.g. random forest, Bayes e.g. Bayes net, lazy e.g. locally weighted learning (LWL), functions e.g. logistic regression (LGR), support vector machine (SVM), etc. (Furxhi et al., 2019a). The application of ML in nanoecotoxicology is used for one of the following reasons, namely; (i) predict the biological (including toxicological) effects of ENPs from their physical, geometric, and chemical properties (these properties are measured experimentally or computed), (ii) screen in silico libraries of virtual nanomaterials and prioritise with the most promising predicted properties, and (iii) guide experimental investigations by focussing costly toxicological studies on a small number of selected and/or rationally designed ENPs (Balraadjsing et al., 2022; Fjodorova et al., 2017; Peng et al., 2020). The most popular ML in Table 2.2 can be ranked as follows: RF, SVM, ANN, MLR, and XGBoost.

2.5.2.1.1 Trees and Rules

Tree-based ML methods construct decision trees to partition a feature space into branches, enabling a hierarchical representation of inputs-output using a similar analogy described in Figure 2.7. In contrast, the rules-based ML approach predicts the

value of a target variable by learning simple decision rules inferred from the data sets (Labouta et al., 2019). Tree- and rule-based ML include RF and DT, respectively (Furxhi et al., 2019a). The DT modelling consists of two steps: generation and pruning. In the first step, a tree of binary split nodes is generated by recursively answering the Yes or No question and evaluating the randomness by measurement of the information gain rate as entropy (Liu et al., 2021).

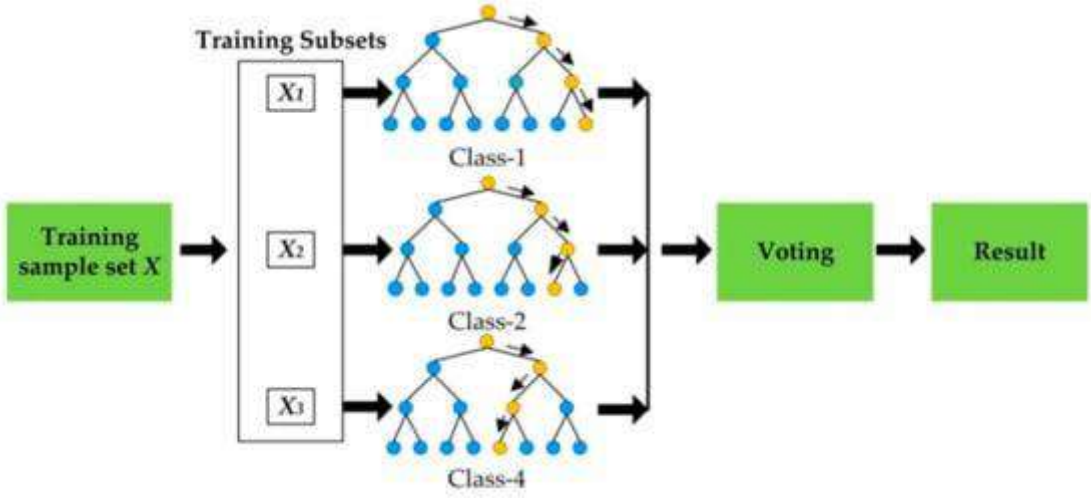


Figure 2. 7. Skeleton diagram depicting the framework of RF. Adapted with permission from (Wu et al., 2019). Copyright © 2019, by the Authors. Used under a Creative Commons (CC BY) License.

RF is an ML model in which a set of random variables determines the behaviour of each tree (Breiman, 2001). RF combined theories such as bagging (Breiman, 1996), random split selection (Dietterich, 1997), the random subspace (Ho, 1998), and the best random split at each node (Amit and Geman, 1997). RF is extensively investigated in most environmental problems, including the nanoecotoxicology domain to probe the fate and toxicity of ENPs (Ban et al., 2018; Duan et al., 2020; Goldberg et al., 2015a). These include examining the influence of parameters of ENPs transport – retained fraction and retention profiles – in saturated columns (Goldberg et al., 2015), screening key parameters that control the reproductive toxicity of ENPs (Ban et al., 2018), and forecasting protein corona on nanomaterials based on novel descriptors (Duan et al., 2020). More other applications are summarised in Table 2.2.

The RF-based models show less overfitting and better performance compared to other advanced ML algorithms (Duan et al., 2020). For example, the RF model exhibited the lowest error and the highest R^2 of 0.78 compared to the least absolute shrinkage and

selection operator (LASSO), and SVM for predicting the antibacterial capacity of NPs (Mirzaei et al., 2021). It was found to be superior to linear discriminant analysis (LDA) for predicting genotoxicity (Ambure et al., 2020) and depicted a stronger predictive ability than DT in estimating the cytotoxicity of photosynthesized nAg (Liu et al., 2021). Other benefits of RF include its independence from the feature input dimension space, ability to rank the relative importance of independent variables, ease of use in interpreting output, ability to handle missing values, and data that contain ambiguities and uncertainties (Duan et al., 2020; Findlay et al., 2018).

2.5.2.1.2 Functions

This class includes the ML algorithm which can be expressed by mathematical functions (Furxhi et al., 2019a). These include SVM, ANN, and MLR. SVM is among the most widely used ML techniques for classification, and regression problems (Cortes and Vapnik, 1995; Vapnik, 1995). Support vector regression (SVR is a component of SVM (Schölkopf and Smola, 2002; Zarei et al., 2018). This discussion just covers the most important topics; for a thorough overview of SVM background theory, see (Cortes and Vapnik, 1995; Smola and Schölkopf, 2004; Vapnik, 1995). SVM provides a high approximation and elucidates the underlying trends in a variety of water-related environmental challenges. These include estimating surface water flow (Granata et al., 2016; Malik et al., 2020), water quality (Mahmoudi et al., 2016; Quan et al., 2020), groundwater (Gong et al., 2016) and probing fate and toxicity of ENPs in nanoecotoxicity (Balraadjsing et al., 2022; Findlay et al., 2018; Furxhi et al., 2019a).

Across a broad range of literature studies, the performance of prediction models constructed using the SVM algorithm typically demonstrates a high accuracy (Gong et al., 2016; Poul et al., 2019; Trinh et al., 2018). For example, the SVM outperformed both the ANN and MLR in predicting the cytotoxicity of $nTiO_2$ and nZnO towards lipid membranes in cells (Papa et al., 2015). In addition, the support vectors exhibited higher accuracy compared to the ANN for estimating the specific heat capacity of aqueous nanofluids of copper oxide (Alade et al., 2019). According to Choubin et al. (2018), the approach's capacity to manage uncertainties through the use of hyperplane is what accounts for the greater prediction. Furthermore, the over-learning issue that frequently occurs with ANN and ANFIS can be resolved by the SVM (Pham et al., 2019).

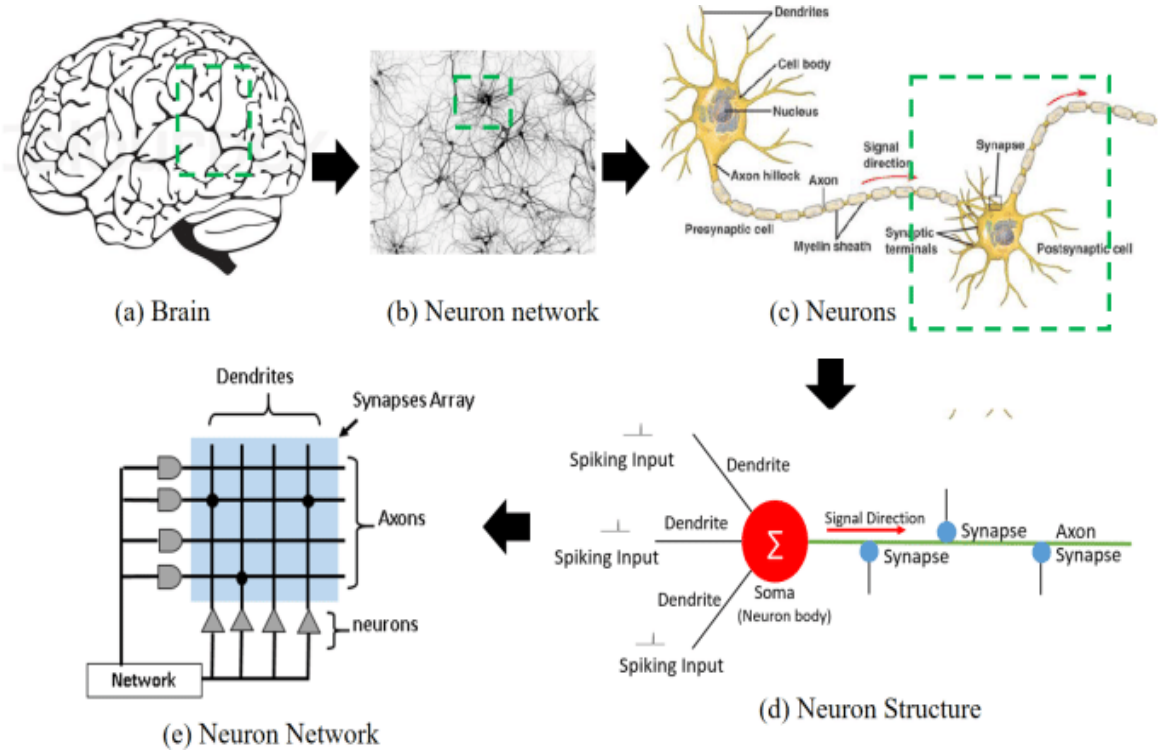


Figure 2. 8. Deep learning networks that emulate the actions shown by real neurons. Adapted with permission from (An et al., 2017). Copyright © 2017, IEEE.

ANNs are potent nonlinear deep learning networks that emulate the actions shown by real neurons as described in Figure 2. 8 (Haykin and Network, 2004; Rosenblatt, 1958). The history and background of ANN are well-established and covered in great detail in numerous studies (Agatonovic-Kustrin and Beresford, 2000; Dharwal and Kaur, 2016; Hornik et al., 1989; Lipton et al., 2015; Müller et al., 1995). Both supervised (labelled) and non-supervised (non-labelled) ANN models have been effectively applied in nanoecotoxicology. Non-supervised models learn on their own and pattern recognition and clustering are the two areas where they are most useful (Zarra et al., 2019). The Kohonen self-organizations (KSO) and counter propagation ANN (CPANN) are two examples of unsupervised models (Fjodorova et al., 2017; Goodner et al., 2001).

Table 2. 2: Research studies using ML in nanoecotoxicology

ML type	*Pre-processing and Validation*	*Findings*		*References*
1. RF *2. LDA*	*1. Randomly split into a 70:30.* *2. 10-fold cross-validation* *3. Genetic algorithm*	*1. Overall statistical prediction quality of the RF model was found to be superior to LDA model.* *2. The LDA model had better sensitivity values (i.e., less false negatives) and thus was better at predicting positives or toxic nanoparticles* *3. RF model had better precision value (i.e., less false positives) and thus was better in predicting negatives or non-toxic nanoparticles.*	*1. Classification*	*Ambure et al, 2020*
1. NN *2. RF* *3. SVM* *4. LG*	*1. Randomly split 70:30* *2. 10-fold cross-validation* *3. vif* *4. SMOTE*	*1. RF and NN with one layer yielded a balanced accuracy of ~97%* *2. All the other developed models had a balanced accuracy of 93%*	*1. Classification*	*Subramanian and Palaniappan, 2021.*
1. RF *2. KNN*	*1. K fold cross-validation* *2. RFE*	*1. RF had a sensitivity of 0.4, a specificity of 0.67, and a balanced accuracy of 0.54 using all input variables.* *2. RF had a sensitivity of 0.8, specificity of 0.83, and balanced accuracy of 0.82 using reduced variables reading.*	*1. Classification*	*Bahl, et al., 2019*
1. KNN *2. NB* *3. RF* *4. ANN* *5. SVM* *6. MLR*	*1. Split into a 60:40* *2. 10 K-fold cross-validation* *3. SMOTE* *4. One-Hot encoding*	*1. Models of RF, kNN and NN had accuracy, precision, sensitivity and balanced accuracy were > 0.7, indicating high predictivity and excellent robustness*	*1. Classification*	*Balraadjsing et al., 2022*

1. *RF* **2.** *SVM* **3.** *LR*	1. *Randomly split into a 90:10* 2. *5-fold cross-validation* 3. *RFE*	1. *RF had the highest score with an F1-score of 0.81 ± 0.02* 2. *The F1-score for SVM and LR algorithms were 0.80 and 0.75, respectively* 3. *A lower F1-score for LR implied that nonlinear relationships are present in the PC data.*	1. Classification	*Findlay et al., 2017*
1. *RF* 2. *NN* 3. *DT* 4. *BN* 5. *SM* 6. *LG*	1. *Split into a 60:40* 2. *One-Hot encoding* 3. *SMOTE* 4. *Copeland Index*	1. *RF and NN showed the best performance in contrast to LWL and IBk which showed the lowest performance.* 2. *LG performance was poor for most metrics and the Copeland Index* 3. *RF had a better RF, performance than NN in some cases*	1. Classification	*Furxhi et al., 2019*
1. *RF*	1. *Randomly split into a 90:10* 2. *10-fold cross-validation* 3. *SMOTE)*	1. *RF model performed better than non-tissue specific models*	1. Classification	*Furxhi et al., 2020*
1. *SVM* 2. *MLR* 3. *RF* 4. *AdaBoost* 5. *NB, LDA*	1. *negative logarithmic scale*	1. *The LDA model appeared to be the best among the 7 tested models.* 2. *The top 3 developed models were MLogitR > AdaBoost > LDA* 3. *The LDA model had 100% sensitivity.*	1. Classification	*Kar et al., 2021*
1. *RF* 2. *KNN* 3. *ANN*	1. *Five-fold cross-validation* 2. *step-wise algorithm*	1. *The predictive ability of the models yielded cross-validated coefficients q2 = 0.58- 0.80 for regression models and balanced accuracies of 65-88% for classification models.*	1. Classification	*Kovalishyn et al., 2017*

1. DT	1. K 10-fold cross-validation	1. DT classifiers were accurate, with high prediction power	1. Classification	Labouta et al., 2019
1. DT 2. RF	1. K 10-fold cross-validation 2. Gini	1. The performance ranking of models from high to low followed the order: RF2 > RF1 > DT2 > DT1. 2. RF2 performed the best with AUC (0.904) and F-measure (0.843) while DT1 performed the worst.	1. Classification	Liu et al., 2021
1. LASSO 2. RR 3. ENR 4. RF 5. SVM	1. Split into a 70:30	1. RF model exhibited the highest R^2 score compared to the other algorithms employed (LASSO, RR, ENR, SVM)	1. Classification	Mirzaei et al., 2021
1. SVM 2. RBFNN	1. Split into a 70:30	1. SVM radial outperformed the other methods and represented a valid alternative to the MLR model.	1. Classification 2. Regression	Papa et al., 2015
1. BPNN 2. RF	1. 10-fold cross-validation	1. RF was better than BPNN, especially for lipid and carbohydrate metabolism 2. The AUC obtained by RF was closer to 1 compared to that obtained by the BPNN	1. Classification	Peng et al., 2020
1. K-MC 2. AdaBoost 3. RF 4. PLS	1. HM 2. GA 3. Split into a 75:25	1. PLS had R^2 and RMSE of 0.38 and 0.18 respectively, and the modelling effect was poor. 2. AdaBoost and RF models had R^2 of 0.78 and 0.85 and RMSE of 0.12 and 0.10. 3. RF model had better prediction ability, robustness and generalization ability.	1. Regression	Sang, et al., 2022

1. RF	1. k-fold cross-validation 2. Split into an 82:18 3. RFE	1. RFR showed satisfactory statistical performance, with the accuracy of the test set and the training set of 0.950 and 0.966, respectively	1. Classification	Shi et al., 2021
1. eXGB 2. LG 3. NB 4. KNN 5. SVM 6. DT 7. RF 8. NN 9. LASSO	1. 10-fold cross-validation 2. Split into a 70:30 3. The LabelEncoder 4. One-hot coding 5. grid-search CV	1. eXGBoost model showed an appreciable prediction R^2 of 0.96	1. Classification 2. Regression	Dong et al., 2022
1. RF	1. 5-fold cross-validation 2. One-hot encoding 3. SMOTE	1. RF showed high prediction with results of R^2 = 0.8530, MSE = 0.0165	1. Classification 2. Regression	Zhou et al., 2021
1. RF 2. SVM	1. Cross-validation (10-fold)	1. RF algorithm and SVR showed high accuracy of greater than 85% and 80, respectively	1. Classification	Trinh et al., 2018
1. RF	1. RF imputation 2. Wilcoxon signed-rank	1. The model demonstrated higher accuracy with R^2 =0.68.	1. Regression	To et al., 2019
1. LR 2. SVR-RBF 3. SVR – Linear 4. RF	1. Split into an 80:20	1. RF and SVR with RBF kernel resulted in high scores of 0.88 and 0.81 in cross-validation, respectively, 2. LR and SVR with linear kernel resulted in low scores	1. Regression	Takahashi et al., 2019

1. *ANN*	1. *Split into a 75:25* 2. *QSTR-perturbation*	1. *ANN showed an accuracy higher than 97% in both training and validation sets.*	1. Classification	*Concu et al., 2017*
1. *eXGBoost*	1. *split into a 70:30* 2. *One-Hot encoding,* 3. *random sampling* 4. *Stratification splitting*	1. *The eXGBoost models for Z-average hydrodynamic size and polydispersity index PDI showed an R^2 of 0.76 and 0.75 for synthesized $nSiO_2$* 2. *For other ENPs, the model achieved R^2 of 0.82 and 0.84 for the Z-average and PDI.*	1. Regression	*Glaubitz et al., 2022*
1. *SVM(Rbf, Linear)* 2. *DT* 3. *RF* 4. *LG* 5. *KNN*	1. *10-fold cross-validation* 2. *SMOTE*	1. *Overall, kNN, RF, Bayesnet, and DT exhibited satisfactory performances in both the training and test sets with high prediction accuracy (97% and 96% in the training and test sets, respectively)*	1. Classification	*Huang et al., 2022*
1. *DT* **2.** *SVM* **3.** *LG*	1. *10-fold cross-validation*	1. *The DT model was the top performer of the three models.* 2. *The ACC values in the training and test set reached 95% and 92%, respectively.* 3. *The AUC reached 95%.*	1. Classification	*Huang et al., 2020*
1. *RF*	1. *Split into an 80:20* 2. *5-fold cross-validation*	1. *RF had the values for precision, recall, and f1 score equal to 0.80 using only protein descriptors.* 2. *The R^2 increased from 0.77 with only protein descriptors to 0.78 by adding the size/charge.*	1. Classification	*Duan et al., 2021*
1. *RF* **2.** *KNN*	1. *Split into an 80:20* 2. *5-fold validation*	1. *KNN models showed better predictability than the RF models*	1. Regression	*Yan, 2019*

		2. The predictabilities of these two models for external validation were similar.		
1. RF	1. Split into an 80:20 2. KruskalWallis (KW) test 3. VIF	1. The RF model yielded high correlations with R^2 of 0.828 - 0.956 between the experimental and simultaneously predicted endpoint toxicity values	1. Regression	Basant and Guta, 2017.
1. PCA 2. SVM 3. RF 4. LR 5. NB	1. 10-fold cross-validation	1. The LR model's mean cross-validation score was equal to 1.0 2. RF modes had a mean cross-validation score of 0.917. 3. LR was a more robust model compared RF 4. Using multiple performance criteria, an RF method was selected as the model with the best performance.	1. Classification	Kotzabasaki, et al., 2020
1. CP ANN	----	1. CP ANN models showed a good prediction power of models with accuracy in the range of 0,83 to 0,92	1. Classification	Fjodorova et al., 2017
1. DT	1. Split into an 80:20	1. The model accuracy for the training and test sets was 100% with small data sets, 2. The accuracy of the model predictions was satisfyingly high and highly statistically significant	1. Classification	Oksel et al., 2016
1. RF 2. LR	1. 10-fold cross-validation	1. LR model performed poorly for the protein corona, as measured by R^2 (less than 0.40). 2. RF had R^2 over 0.75 in the prediction	1. Regression	Ban, et al., 2020
1. RF	1. Split into a 60:40 2. Data gap filling	1. RF showed the precision and accuracy were greater than 80%	1. Classification	Ha et al., 2018

1. GL 2. SVM 3. RF 4. NN 5. KNN	1. 10 fold cross validation; 2. SMOTE 3. one hot coding 4. Normalization z-score, 5. min-max, log10	1. The NN model built using the balanced data set was identified as the model with the best predictive performance. 2. RF showed roughly similar performance regardless of the normalization method used.	1. Classification	Choi et al., 2018
1. BPNN	1. Split into a 70:30 2. GEP	1. The BPNN model showed an R^2 higher than 0.8 for all simulations.	1. Classification	Wang, et al 2021
1. RF 2. LR	1. curve-fitting	1. The RF achieves excellent classification accuracy of greater than 95% and small error rates of less than 4%).	1. Classification 2. Regression	Ban et al., 2018
1. RF	1. 5-fold cross-validation 2. RFECV 3. Split into a 90:10	1. Regression results demonstrated that the fraction of nanoparticle mass retained over the column length can be predicted with an expected mean squared error between 0.025–0.033.	1. Classification 2. Regression	Goldberg et al., 2015

Synthetic Minority Oversampling Technique (SMOTE), recursive feature elimination (RFE), Genetic algorithm GA, variance inflation factor VIF, gene-expression programming (GEP), *partial least squares (PLS), heuristic method (HM), Area under the AUC), Radial basis function neural networks (RBFNN); linear discriminant analysis (LDA); Generalized linear model (GL); back propagation neural network (BPNN); Multi-Linear Regression (MLR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF) or k Nearest Neighbors (kNN). genetic programming-based decision tree GPDT; Linear Regression, LG: Generalized Linear Regression; GLR; Support Vector Machines, eXtreme Gradient Boosting; GB; Logistic Regression, "LR" and Naive Bayes, "NB"); (Principal Component Analysis, PCA); gradient boosted decision tree GBDT, Least Absolute Shrinkage and Selection Operator (LASSO) Regression, Ridge Regression (RR), Elastic Net Regression (ENR), SVM multinomial (elastic net); Least-squares linear regression; least absolute shrinkage and selection operator (LASSO),*

On the other hand, the supervised ANN models are the widely applied approach and permit adjustment of the weights and bias relationship to generate the required output accuracy for regression and classification problems. The feed-forward neural network (FFNN), associative neural network, convolution neural network (CNN), and feedback/recurrent network (RNN), are examples of supervised models. Associative NN is the combination of MLP and K-nearest neighbor (KNN) (Kovalishyn et al., 2018). The FFNN mechanism proceeds in one direction, from the input to the output nodes (Dharwal and Kaur, 2016; Papa et al., 2015). RNNs generate a closed cycle with feedback connections to the preceding nodes and are particularly difficult to train (Lipton et al., 2015; Lukoševičius and Jaeger, 2009). FFNN includes multilayer perceptron (MLP), single-layer perceptron (SLP), and radial basis function neural networks (RBFNN). SLP method is typically ineffective for handling complex issues (Pradhan et al., 2020). The RBFNN and MLP networks are the FFNN framework types that are most frequently used in numerous fields. RBFNN is made up of the radial basis activation function (RBF) in the hidden layer and requires centers equal to the dataset entries (Valizadeh and Sohrabi, 2018).

ANNs are quite useful and have a rich history demonstrated, in modelling dynamic nonlinear systems, where traditional statistical methods might not provide reasonable approximations (Alimissis et al., 2018; Hou et al., 2020; Pradhan et al., 2020). Many environmental-related fields application over the last decade include hydrological systems (Cavalcante et al., 2013; Pradhan et al., 2020), and water quality assessment (Cordoba et al., 2014; Sarkar and Pandey, 2015), river sediments (Olyaie et al., 2015; Rajaee et al., 2009), simulation of heavy metals in aqueous systems (Lu et al., 2019), evaluating particle size distribution (Lagos-Avid and Bonilla, 2017), determine the runoff (Jimeno-Sáez et al., 2018; Kumar et al., 2016), microbial ecology (Larsen et al., 2015; Santos et al., 2014).

Theoretical assessments of the fate and toxicity of ENPs in nanotoxicology to various endpoints using ANN have been summarised in Table 2.2. These assessments include investigating cytotoxicity to Escherichia coli (E. Coli) (Fjodorova et al., 2017), immobilisation of Daphnia magna (Balraadjsing et al., 2022), metabolic pathways (Peng et al., 2020), and cytotoxicity of ENPs based on cell viability (Furxhi et al., 2019a), among other endpoints. The coefficient of the determination (R^2) was 0.8 for predicting the uptake and translocation of NPs in plants (Wang et al., 2021). ANN

demonstrated an average accuracy of 0.83 to 0.92 in predicting the cytotoxicity of metal oxide ENPs (size between 15 and 90 nm) to the bacteria E. Coli (Fjodorova et al., 2017). ANN showed an accuracy of 81 ± 7.0% for estimating the MIC (MIC- is the lowest concentration of the toxicant needed to produce an inhibitory effect) of metallic nanoparticles (Kovalishyn et al., 2018).

ANN demonstrates good generalisation and prediction accuracy, comparable to RF but relatively high to the other algorithms, such as K-nearest neighbors, etc. (Balraadjsing et al., 2022; Subramanian et al., 2021; Furxhi et al., 2019; Choi et al., 2018; Peng et al., 2020). Apart from the impressive prediction performance seen in these investigations, ANN also offered a more profound understanding of the physicochemical characteristics or variables in charge of predicting the toxicity of ENPs (Concu et al., 2017). For example, the four most crucial factors for predicting the toxicity of ENPs were found to be hydrodynamic size, exposure time, formation enthalpy, and dose in an analysis of relative attribute importance using the ANN model (Choi et al., 2018).

Furthermore, linear methods including multi-linear regression (MLR) or logistic regression (LG) models are generally less accurate when considering complex nonlinear relationships between inputs and outputs, leading to inconsistencies in the intended response (Aquilina et al., 2018). MLR and LG are hinged on pre-defined basic relationships between predictors and output variables, in which, under the circumstance where data have no fundamental correlation, like the case of ENP nanoecotoxicology data, the model can yield unsatisfactory results (Chen et al., 2019; Zhang et al., 2018). Thus, these models perform comparably to other high-performing ML models such as ANN and RF for elucidating linear systems (Lee et al., 2016).

2.5.2.1.3 Hybrid system

ML of ANN, SVM, and RF among others, has been demonstrated as successful in prediction. However, the application of hybrid systems over the years has attracted attention. Hybrid ML models combine characteristics of two or many ML techniques. Examples of these models include XGBoost or ANFIS, etc. In this case, despite RF being the ensemble method of DTs, it was not considered to be a hybrid system, since it is solely based on combination DTs. XGBoost is based on an arbitrary differential loss function and the gradient descent optimization procedure. This gives the technique its name gradient boosting as the loss gradient is minimized as the model

fits much like ANN. XGBoost offers good efficiency and flexibility while being relatively fast and relatively easy to implement as well as interpret (Chen et al., 2015). XGBoost models are optimal for limited datasets due to their robustness in comparison with e.g., a deep-learner. Additionally, they enable a straightforward ranking of the feature importance (Dong et al., 2022; Glaubitz et al., 2022).

XGBoost and gradient boosting (GB) use the same principle, but the former has regularisation to reduce overfitting and bias (Ogunleye and Wang, 2019). Regularization pushes the weights for many variables to zero and thus performs variable selection, which plays a key role in high-dimensional problems. When the regularization parameter is set to zero, the objective reverts to the traditional gradient tree boosting or GBM (Chen & Benesty, 2016). As the results of a bias-variance trade-off of 1, the XGBoost derived generally shows high prediction accuracy. XGBoost showed an appreciable prediction R^2 of 0.96 for assessing comparable bioconcentration potentials for nanoparticles in aquatic organisms (Dong et al., 2022) In addition, the XGBoost showed an R^2 of 0.76 and 0.75 estimating the Z-average hydrodynamic size and polydispersity index (PDI) of synthesized nSiO_2 (Glaubitz et al., 2022).

The concept of a neuro-fuzzy inference system was introduced by Jang, (1993). An ANN and FL principles are combined in a neuro-fuzzy inference system (Jang and Sun, 1995; Jang, 1993). ANN models lack interpretable features despite having great computational and prediction skills. FL, on the other hand, lacks systematic methods for determining the MF parameters and has high interpretable attributes (Castillo and Melin, 2014; Zimmermann, 2010, 2011a). Therefore, in a neuro-fuzzy inference system, the if-then rules and MF that are challenging in FL are generated using the learning capability of ANN, in turn, the FL enhances the interpretability attribute that is lacking in ANN (Gong et al., 2016; Zaghloul et al., 2020).

Adaptive network-based fuzzy inference system (ANFIS) is the most popular and simplest form of neuron-fuzzy systems (Jang, 1993). ANFIS generally shows a profound ability to extract fuzzy rules from arithmetic data to improve interpretation. It generates if-then that are necessary to aid in decision-making while also simplifying the complex interaction among variables (Coutinho et al., 2019; Pham et al., 2019). Other modified forms are self-organizing fuzzy neural networks (SOFNN) (Leng et al., 2005), dynamic fuzzy neural networks (DFNN) (Wu et al., 2000), generalized dynamic

fuzzy neural networks (GDFNN) (Zhu et al., 2007) and dynamic parsimonious fuzzy neural network (DPFNN) (Pratama et al., 2013). Excellent applications of neuro-fuzzy systems have been used for estimation and optimization of the water treatment disinfectant process (Hong et al., 2018), surface water flow (Belvederesi et al., 2020; Mohammadi et al., 2020), groundwater and hydrology systems (Emamgholizadeh et al., 2014; Gong et al., 2016), among other environmental domains. However, there is a lack of a body of information regarding the application of neuro-fuzzy systems in nanoecotoxicology.

2.5.2.2 Knowledge-based systems

A KBS is a computer program that leverages a centralized information repository to support decision-making. KBSs are an alternative to traditional data-driven methods like ML. ML systems learn from data without explicit programming. Conversely, KBSs rely on a reasoning system to derive new knowledge (Amirshenava and Osanloo, 2019; Grella et al., 2019; Obiedat and Samarasinghe, 2016). This makes them a valuable tool for solving complex problems across a wide range of applications. They are capable of processing expert insight by combining judgment and heuristics (Obiedat and Samarasinghe, 2016).

Figure 2.9 shows a typical KBS structure. The reasoning system, or intelligent processing, which generates inference to arrive at a certain solution, is the essential component of KBS. To improve decision in the inference many KBS use mathematical theories like fuzzy set theory and Bayesian network (Money et al., 2012) or semi-quantitative analysis (Tiede et al., 2015). For example, Money et al. (2012) developed a Bayesian network integrated with expert knowledge for estimating the nAg concentrations in aquatic environments.

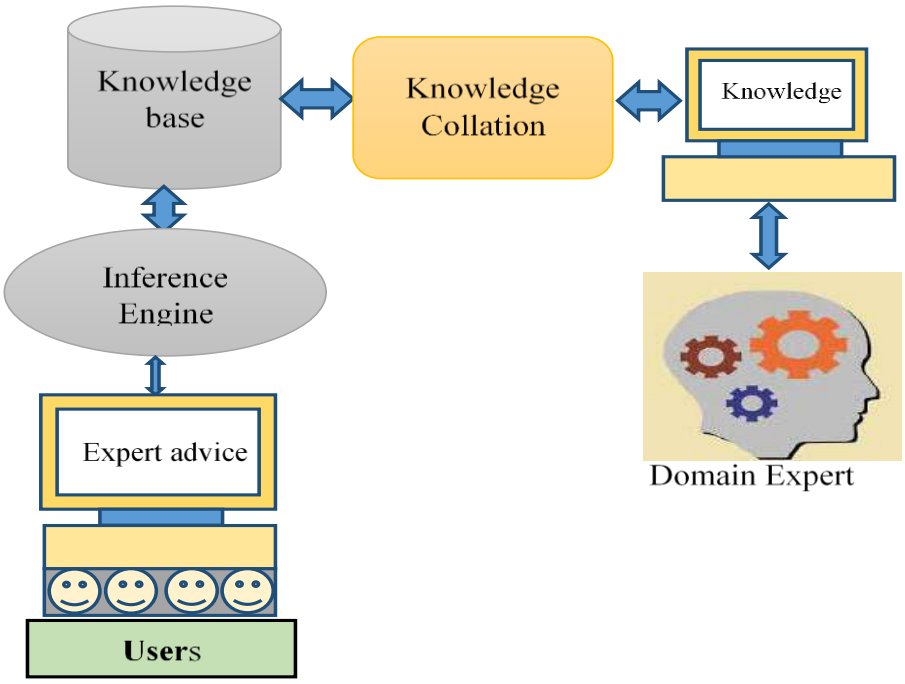


Figure 2. 9. General structure for knowledge-based systems (Gennari et al., 2003).

2.5.2.2.1 Weight-base inference

The use of ranking and scoring factors, rather than the application of mathematical terms used in most mechanistic models is known as semi-quantitative analysis (SQA) (Giubilato et al., 2014; Narita et al., 2014). Scores express the weight of the impact of each factor based on expert judgment and heuristics. The semi-quantitative model (SQM) is similar to qualitative examinations; as testing does not measure the precise quantity of a substance. However, the SQA uses numerical scores or indices, whereas qualitative modelling uses descriptive or linguistics (Pang and Coghill, 2015a).

By reducing personal bias during the evaluation process, the implementation of these rating or scoring systems in SQM ensures consistency and reproducibility of the results (Obiedat and Samarasinghe, 2016; Tang et al., 2019). The application of the SQA approach is found in multi-criteria decision analysis (MCDA) (Giubilato et al., 2014; Narita et al., 2014; Singer et al., 2017), analytic hierarchy process (Balsara et al., 2019), analytic network process (Zhang et al., 2016), and simple multi-attribute rating technique (Siregar et al., 2017). Multi-criteria decision-making (MCDM) offers a useful framework for processes that prioritize different variables according to a range

of criteria. To rate factors, the widely applied scoring methodologies are Saaty's (Saaty, 1980) and American Chemistry Council ranking systems (ACC, 2011).

SQM performs well in handling environmental issues in general (Pilone and Demichela, 2018; Singer et al., 2017) and there hasn't been much research published in the field of nanoecotoxicology. Musee (2017) applied a semi-quantitative-based model to rank the potential risks of consumer nano-products. Additionally, Tiede et al. (2015) used the approach to estimate the exposure of ENPs in drinking water for humans. In the respective studies, the approach yielded good risk characterisation, was less dependent on the quantitative data and the number of assumptions are few relative to pure quantitative or qualitative methods (Grella et al., 2019; Tang et al., 2019). The capacity to handle various imprecise numerical values, incorporate uncertainty resulting from the modelling of expert intuitions or perceptions using scores or weights, and oversimplification are some of the disadvantages (Ye et al., 2020).

2.5.2.2.2 Rule-based inference

Expert systems (ES) are the most well-known type of KBS. ES uses qualitative modelling in the format of conditional statements of the rules 'if' (satisfied set of conditions) and "*then"* (a set of consequences can be inferred) (Patterson, 1990). ESs have challenges in dealing with uncertainty and are time-consuming; thus to address the uncertainty and ambiguity of ES or modelling using weights a more compatible and efficient approach is FL. FL is founded on the concept of fuzzy set theory (Zadeh, 1965), which is an extension of classical sets (Brubaker, 1985).

In classical sets, the true value of a proposition only takes the crisp values 0 (non-membership) or 1 (total membership). This means for a set A that is defined in the universe of discourse R in the ranged [0, 1], $A \in R$ only if $\mu_A(x) = 1$, otherwise for $\mu_A(x) = 0$, $A \notin R$. On the other hand, FL is a multivalued procedure, in which sets have an infinite number of degrees of membership described by fuzzy sets and MFs values with the universe of discourse [0,1] (Araya-Muñoz et al., 2017; Mazhar et al., 2019; Musee et al., 2006). This mathematical function allows sets to overlap and the ability to integrate uncertainty.

There are two types of fuzzy sets: types 1 and 2. Equation 2.1 describes type-1 fuzzy sets, which are the conventional single MF. The type-2 fuzzy sets generalize type-1

fuzzy sets using double membership functions (Jana, 2016; Roy et al., 2019). The simplest type, type-1, is described in this work. The background and fundamental principles of fuzzy logic-driven modelling are well-known and extensively covered in great detail (Mendel, 1995; Zadeh, 2008; Zimmermann, 2011b), therefore, for convenience, herein will not be provided.

$$A = \{x, \mu_A(x) | \, x \epsilon R\}, \quad \mu_A(x) x \epsilon \, [0,1] \tag{2.1}$$

Where A is the fuzzy set described using linguistic variables and $\mu_A(x)$ is a value of an MF in the interval [0, 1].

To date, several studies have proposed and reported the use of FL to address the risk of ENPs based on case studies to demonstrate application (Ramirez et al., 2022; Topuz and van Gestel, 2016). For instance, nAg and $nTiO_2$ were used as case studies to demonstrate the suggested analytical hierarchy method (AHP) to assess the risks of ENP in various environmental compartments (Topuz and van Gestel, 2016). In these respective studies, the FL provided inherent features that allowed the manipulation of data, the flexibility in handling imprecise/vagueness, easy-to-understand and interpretation, and less dependence on large data.

2.5.3 Advantages and disadvantages

The application of ML and KBS has been increasingly popular in recent years. While these techniques are beneficial, they share both advantages as well as disadvantages as listed in Table 2.3. KBS uses explicitly programmed rules, while ML systems learn from data without explicit programming. KBS tend to be more transparent but less adaptable, whereas ML systems are adaptable but may lack transparency in their decision-making processes.

Table 2. 3. Advantages and disadvantages of algorithms.

Method	Advantages	Disadvantages
MLR	1. Simple 2. Transparent 3. Easy to interpret	1. Nonlinear relationships cannot be detected 2. Needs to follow the “Topliss and Costello rule” 3. Easily affected by outliers
ANN	1. Allows nonlinear relationships to be revealed 2. Can manage uncorrelated and/or uncertain data 3. Can be used for large or small data sets	1. High sensitivity to changes in the model parameters 2. Prone to overfitting 3. Difficult to interpret 4. More computation time
DT	1. Transparency 2. Easy to interpret 3. Small and large datasets are both suitable 4. Allows nonlinear relationships to be revealed 5. Able to handle both numerical and categorical data	Classes cannot overlap 2. Predictive ability is low 3. Prone to overfitting
SVM	1. Allows nonlinear relationships to be established 2. Can deal with a small dataset 3. Low risk of overfitting	1. Difficult to interpret 2. High sensitivity to changes in the model parameters
RF	1. Allows to reveal nonlinear relationships 2. Relative high accuracy	1. Need to prune the tree 2. Imbalanced data affects the results 3. Small datasets
kNN	1. Can be used to reduce the number of features 2. Low risk of overfitting 3. Easy visualization	1. Difficult to interpret the meaning of independent variables
XGboost	1. High accuracy: 2. Speed 3. Regularization 4. Flexibility	1. Complexity 2. Overfitting 3. Memory usage 4. Lack of transparency
FL and SQA	1. Results easy to interpret 2. Integrate ambiguity and impression 3. Expert knowledge	1. Extensive prior knowledge required 2. Structure/relationship known 3. Huge effort to develop prior knowledge 4. Effort required to gather knowledge

2.6 Chapter summary and knowledge gap

This chapter discussed the progressive application, possible emission pathways, environmental concentration, and application of several ML and KBS as an efficient cost-effective means of addressing the exposure and risk assessment of ENPs in nanoecotoxicology domain. Across a broad range of studies reviewed the following knowledge gaps were identified: i) despite the dissolution and aggregation 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 transformation in aquatic systems, ii) despite the great amount of work reported to elucidate the exposure and risks of ENPs, cost-effective and easy-to-use computational exposure characterisation tools that do not require software are lacking and scarce, and iii) suitable approach that can support the reasoning of well-thought-out solutions to complex problems that are characterised by uncertainty, poorly defined information, lack of sharp boundaries and quantitative data are also lacking.

Chapter 3. Materials and Methods

This chapter provides a detailed discussion of the data collecting process, as well as methods for training and evaluation of proposed KBS and ML models. ML primarily deals with training on data to learn patterns and make predictions. This data can be labelled (supervised learning) or unlabelled (unsupervised learning), and consists of input-output pairs or features and labels. On the other hand, KBS uses structured data, including rules, facts, and relationships. The data in KBS is explicitly represented and encoded by human experts to provide a basis for reasoning.

3.1 Meta-data analysis and systematic review

The field of nanoecotoxicology has seen a steady increase in experimental data recently; yet, there is a lack of accessible nano databases e.g. NanoE-Tox (Juganson et al., 2015), and S2NANO database (www.s2nano.org) (Trinh et al., 2018) for efficient analysis. The meta-data analysis and systematic review (MDASR) process which makes use of secondary data extracted from literature reviews was considered (Foley et al., 2018; Gurevitch, 1993; Haidich, 2010). MDASR has a long history in several environmental fields, including nanoecotoxicology research (Ban et al., 2018; Furxhi et al., 2019a; Goldberg et al., 2015a) and exposure of microplastics in aquatic systems (Sun et al., 2021; Foley et al., 2018; Spear et al., 2016).

The benefits of MDASR are the capacity to combine pertinent individual research to capture all conceivable permutations that underlie the phenomenon under study, and the reduction of noise and bias present in individual studies (Deji et al., 2021; Greco et al., 2013; Gurevitch et al., 2018). The application of MDASR adhered to the guidelines of "Preferred Reporting Items for Systematic Reviews and Meta-analyses" (PRISMA) (Moher et al., 2009). The process constituted the following key steps, namely (i) developing the search strategy and defined boundary for rapid or systematic search, (ii) screen studies and determine eligibility, (iii) quality assessment of studies, and appraisal (Tolaymat et al., 2017).

3.1.1 Search strategy

Online databases such as PubMed, Google Scholar, ScienceDirect, American Chemical Society, SpringerLink, and Web of Science were used to search peer-reviewed scientific publications. Key terms like "fate OR deposition" AND "aggregation

OR hetero-agglomeration OR homo-aggregation OR the surface transformation" AND "nanoparticles OR nanomaterial" AND "aqueous media OR natural water OR freshwater" AND "fate OR adsorption OR dissolution OR dispersion" were used singly and/or in combination (using a set of Boolean logic operators like AND, OR, and NOT) during the search. Figure 3.1 displays the density visualisation map of key terms generated using the VOSviewer program (https://www.vosviewer.com/).

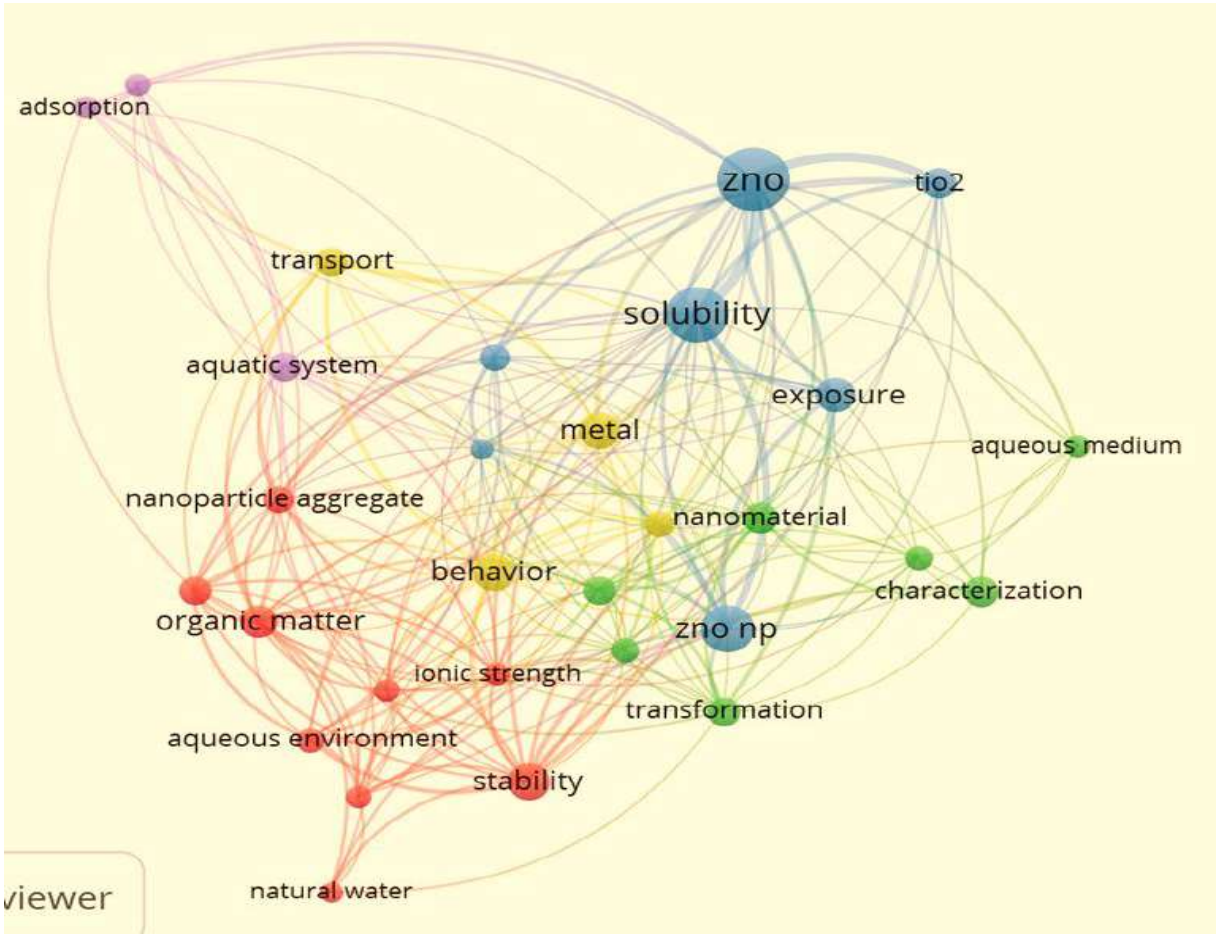


Figure 3. 1. Density visualisation map showing co-occurrence of the keywords generated using the VOSviewer program.

3.1.2 Criteria for inclusion and exclusion of research studies

The peer-reviewed articles were screened using the following criteria (i) metal/metal-oxide as ENPs of interest, (ii) freshwater-like (e.g. river, lakes, synthetic media, etc.) exposure media and (iii) the identified studies needed to provide key information on the physical-chemical properties and water chemistry properties. These include shape, surface area, size, morphology, coating, ZP, solubility, and water chemistry

factors such as NOM (dissolve organic carbon (DOC), total dissolve carbon (TOC), IS, pH, conductivity, and temperature, among others (Tolaymat et al., 2015). Research studies on soil, wastewater, and marine environment transformations were excluded from consideration. The bibliographies of retrieved articles were searched for pertinent references not found in the electronic search. Only documents and reports written in English were taken into account.

3.2 Data

3.2.1 Extraction of data

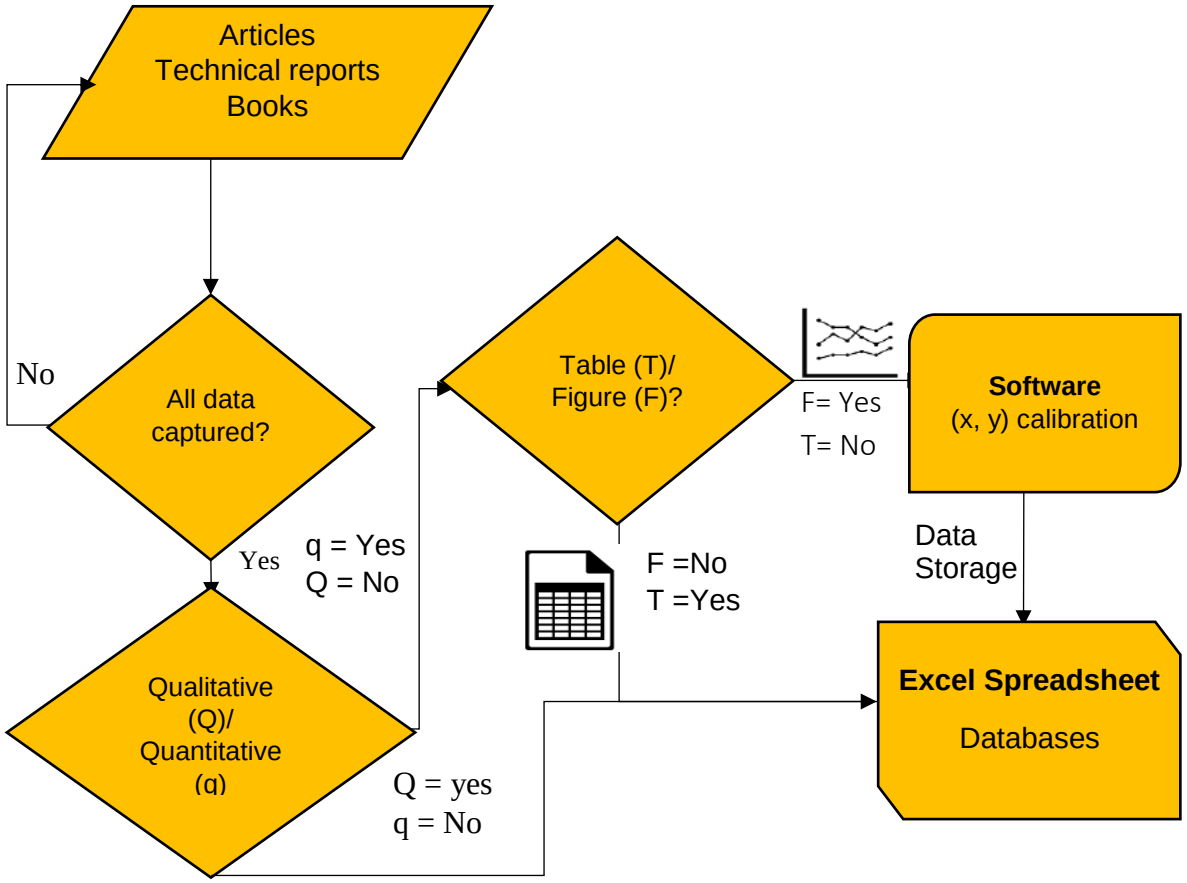


Figure 3. 2. Framework to develop a database based on the use of literature sources reported on aqueous systems (Gagliardi et al., 2016).

Figure 3. 2 depicts the framework followed for extracting both qualitative and quantitative data from literature sources. Extracting quantitative data from research papers where, for example, data embedded in figures can be achieved using several methods. These consist of an online program called GetData Graph Digitizer, a manual method using expanded graphs (Church, 2002), and a MetaLab module (Mikolajewicz and Komarova, 2019). According to Kadic et al. (2016), the manual

method is laborious, time-consuming, and prone to human mistakes. On the other hand, there have only been a few recorded applications of MetaLab efficiency. GetData Graph Digitizer v.2.24 software (http://getdata-graph-digitizer.com/) (Kadic et al., 2016; Wang et al., 2020) which has been applied in numerous research studies (Gagliardi et al., 2016; Sun et al., 2021; Wang et al., 2017; Yang et al., 2020) was applied.

3.2.2 Data quality and cleaning

Maintaining high-quality data is essential to gain valuable insights, perform accurate analyses, and make well-informed decisions. The quality of data gathered can be influenced by various factors, such as data collection methods, data entry processes, data storage, and data integration. In particular, integrating and standardizing data from many sources and formats can be difficult, resulting in inconsistencies and inaccuracies (Balraadjsing et al., 2022; Glaubitz et al., 2022). In this work, overcoming the difficulties of integrating data from different sources and ensuring the data was of high quality, several approaches were followed such as data cleansing, data quality profiling, and data standardization. Data cleansing examines and cleans the data for mistakes, duplicate records, and inconsistencies. Duplicate entries result in bias analysis by over-representing specific data points or trends. Therefore redundant records or duplicates were eliminated to improve the uniqueness of the dataset (Findlay et al., 2018; Furxhi et al., 2019a).

Furthermore, data quality profiling was undertaken to analyze existing data and summarize it. Data profiling aids in the identification of remedial measures to be performed and generates important insights for improvement initiatives. As the results of the meta-analysis adopted in this study, profiling of existing data was undertaken consistently to ensure data gathered were of quality at all times. Lastly, secondary data gathered are generally defined as inconsistent data formats and units of measurement for data entry. To ensure consistency, the environmental indicators—which were primarily measured using a variety of scales and units—were harmonized into a single standard metric system through a data standardisation process. This step is essential for eliminating inconsistencies and inaccuracies. For example, the measurement of input such as NOM was changed from mol/ ℓ to mg/ ℓ. In research studies that reported the concentration of ions (Na^+, Cl^-, HCO_3^- Ca^{2+}, SO_4^{2-}, Mg^{2+}, SO_4^{2-}, and K^+), the Deybe-Huckel formula in Equation 3.1 was utilized to compute IS.

$$I = 0.5 \sum_{i}^{n} c_i \, z_i{}^2 \tag{3.1}$$

Where *I* is the ionic strength, *n* is the total number of ionic species, *i* represents the specific ionic speciation, c_i is the molarity of the ionic species, and z_i is the charge on the ionic species (French et al., 2009).

3.3 Identification of model input and output (s)

Multiple factors influence the fate and behaviour of ENPs. The number of inputs-outputs required for modelling was chosen using the evidence-based procedures (EBP) (Akobeng, 2005), and Occam's Razor parsimonious (Blumer et al., 1987) concepts. The EBP is the research practice mostly applied in medicine (Akobeng, 2005; Sackett et al., 1996) for problems defined by scarcity of well-curated data. In recent times increasing applications have also been witnessed in nano-ecotoxicology (Kozleski, 2017; Tolaymat et al., 2017, 2015). EBP pre-supposes that the identification of model input and output (s) must be based on the weight of evidence (WOE) and strength of evidence (SOE) gathered in research studies. On the other hand, the parsimonious principle pre-supposes that few parameters can adequately suffice for building a simplified model that is easily comprehensible by policy- and decision-policy makers (Huang and Jolliet, 2016; Pratama et al., 2013).

3.4 Models development

Data is a "commodity" necessary for both ML and KBS: however, the nature of data, representation, and learning approach, differ as discussed in Section 2.5.2. ML learns patterns from data and/or is supervised on structured numerical input-output pairs or features. On the other hand, KBS are explicitly coded with expert knowledge for decision-making. Sections 3.4.1 and 3.4.2 provide details on training and testing of the ML algorithm and development of KBS, respectively.

3.4.1 Developing ML algorithms

The ML predictive algorithms are generally developed from continuous numerical data pairs$(x_1, y_1, \ldots, x_\mathrm{i}, y_\mathrm{i},)$, where input vectors $\mathrm{x} = (x_1, x_2, \ldots, X_\mathrm{i})^\mathrm{J}$ (i is the number of inputs), y_i is the corresponding output, and the value J =1, 2…*n*-dimensional space.

The development of the ML algorithm followed the schematic diagram depicted in Figure 3.3.

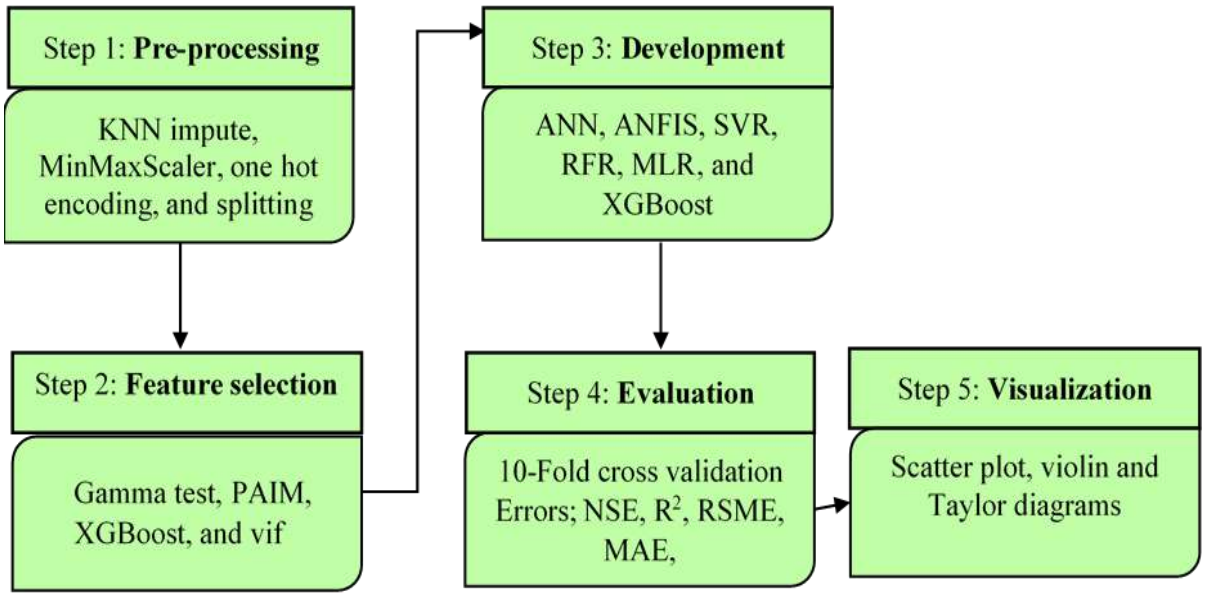


Figure 3. 3. Schematic diagram showing the procedure for training and testing of ML algorithms (Dong et al., 2022).

3.4.1.1 Pre-processing and handling data uncertainties

In data modelling, uncertainty or data veracity is one of the defining characteristics of data. Uncertain data is data that contains noise that makes it deviate from the correct, intended, or original values (Batista and Monard, 2003; Troyanskaya et al., 2001). The uncertainty in the data can arise from multiple sources, including measurement errors, discrete sampling of the measurements, use of both quantitative and categorical variables, use of both structured and unstructured data, and the presence of missing data. To derive decisions, data modelling and analyses must necessarily account for many different kinds of uncertainty present in very large amounts of data. This is because analyses based on uncertain data will affect the quality of subsequent decisions (Lin and Tsai, 2020).

Handling the uncertainty in the data raises challenges in almost all aspects of data management. The uncertainty of data sets was determined from descriptive statistics such as computing standard deviation, mean, and probability distribution of inputs and output parameters. A probability distribution function provides the possible values that

a variable can have and assigns a probability of occurrence to each. Among other approaches, the uncertainty was reduced through data smoothing and increasing the size of a data sample, but there was always some random variability based on the non-linearity of the data under consideration.

3.4.1.1.1 Normalisation

High data variability and large standard deviation (SD) can reduce the sensitivity of the developed ML models. Thus, to avoid oversaturation, the data were normalised and scaled to a mean of 0 and a variance of 1, or between 0 and 1 in Python v.3.0 using the expression in Equations 3.2 and 3.3, respectively:

$$x_{norm,i} = \frac{x_i - \mu}{\sigma} \tag{3.2}$$

Where $x_{norm,\,i}$ is the i^{th} normalized data point, x_i is the i^{th} data point, μ and σ are the mean and standard deviation, respectively. Each data point in the set was normalized based on their respective explanatory variable's mean and standard deviation.

$$\mathrm{y_j}(k) = \frac{\mathrm{x_j}(k) - min\mathrm{x_i}(k)}{max\ \mathrm{x_i}(k) - min\mathrm{x_i}(k)} \tag{3.3}$$

where, j= 1,2…m; k=1,2…n, i=1,2,…,p $\mathrm{x_i}(k)$ is the original sequence, $\mathrm{y_j}(k)$ is the sequence after data pre-processing, $min\mathrm{x_i}(k)$ and $max\ \mathrm{x_i}(k)$ are the smallest and the largest values of $\mathrm{x_i}(k)$, respectively, and $\mathrm{x_j}(k)$ is the data point normalised.

3.4.1.1.2 One hot encoding

Categorical features cannot be utilised by ML algorithms and thus were converted to numerical data before modelling (Findlay et al., 2018; Furxhi et al., 2019a). Categorical variables were converted by applying one hot encoding following the pseudocode described in **Algorithm 1**. The first step involved the identification of features with categorical variables. Based on the uniqueness of the categories one hot encoding was used to create multiple binary numerical features or dummy variables using 1 or 0 to indicate the presence or absence of the feature (Balraadjsing et al., 2022; Glaubitz et al., 2022). The resultant data sets had increased dimensions.

Algorithm 1 One hot encoding

Procedure:

1. Identify the features with categorical variables
   - features = {}
2. Identify unique categories:
   - $unique_{categories}$ = unique (features)
4. Initialize a matrix to store the one-hot encoded representation:
   - one_hot_encoded_data = []
5. Assign dummy variables to each unique category in the new column
   - Presence [1] and absence [0]
6. Save the new data with increased dimension

3.4.1.1.3 Missing data imputation

Highly diverse experimental studies with inconsistent reporting protocols result in heterogeneous data (Basei et al., 2019a). ML techniques generally cannot handle missing information and incomplete data (Lin and Tsai, 2020). MD is categorised as missing completely at random (MCAR): missing data is not related to any other variables; missing at random (MIR): missing data that are related to other variables and not missing at random (NMAR): missing data are related to the variable itself (Batista and Monard, 2003; Troyanskaya et al., 2001). Various strategies (e.g., mode/mean, regression, etc.) exist to deal with missing information. However, despite the benefits linked with the mean imputation approach, replacing all missing records with a single value results in distortion of the input data distribution and statistical bias (Batista and Monard, 2003; Troyanskaya et al., 2001).

In this work, missing values were imputed using the K-nearest neighbors (K-NN) approach described using pseudocode in **Algorithm 2**. The first step involved identifying the missing values. The Euclidean distance was used to quantify the distance between neighbors (Ling and Dong-Mei, 2009). The number of neighbours (k) investigated was from 1 to 20 and p was 2. The k neighbor points that have the shortest distance to the unknown point were used to estimate the missing value (Batista and Monard, 2002; García-Laencina et al., 2009; Pan and Li, 2010). The higher the value of k, the better the performance of the model, while a smaller value of k leads to overfitting. However, there are no predefined statistical methods to find

the most optimum value of k, it is sample size, and dimension dependent (Zhang et al., 2017). The K-NN was applied in Scikit-Learn Library in Python 3.0.

Algorithm 2 K-nearest neighbors (K-NN) impute

Parameters:

- Missing values (M_{values})
- Number of neighbors (k)
- Distance metric
- $x = (x_1, x_2, x_3 \ldots, x_n)$
- $y = (y_1, y_2, y_3 \ldots, y_n)$
- n is the vector size.
- w_i in weight of every single neighbor point y_i
- d_i is the distance between the unknown point and its neighbour

Procedure:

1. Identify features with M_{values}
2. Calculate the distance from neighbors
   - Euclidean distance

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

2.2. Select the k-nearest neighbors:

- Identify the k rows with the smallest distances.

$$y_i = \sum_{i=1}^{n} w_i y_i$$

$$d(x/h) = \frac{d_i}{\sum_{j=1}^{n} d_i}$$

2.3. Impute missing values:

- Replace the missing value with the mean or median of the corresponding feature among the nearest neighbors.

3. Repeat the process for all rows with missing values in the dataset.
4. Save the imputed dataset

3.4.1.1.4 Multicollinearity

Multicollinearity is a statistical concept in which independent variables may show high correlations with each other (Alin, 2010; Subramanian and Palaniappan, 2021). Highly correlated explanatory variables can reduce interpretability, and increase the variance (Subramanian and Palaniappan, 2021). To detect multicollinearity a metric known as the variance inflation factor (VIF), was applied. This approach measures the correlation and strength of correlation between the explanatory variables in a regression model (Mansfield and Helms, 1982). The variance inflation factor (vif) was computed using Equation 3.4.

$$\mathrm{vif_i} = \frac{1}{1 - \mathrm{R_i}^2} \tag{3.4}$$

Where the $\mathrm{R_i}^2$ is the coefficient of multiple determination of x_i (i = 1,2, …, k) on the remaining explanatory variables. The vif score of 1.0 and < 5.0 indicate no correlation, and moderate multicollinearity that does not pose a serious problem, respectively. The vif scores between 5 and 10 and scores> 10 indicated high, and severe multicollinearity, respectively.

3.4.1.1.5 Feature selection (FS)

Feature selection (FS) is the process of gaining a deeper understanding of the preponderant factors that influence the studied phenomena (Chang et al., 2014; Hou et al., 2020; Remesan and Mathew, 2015). FS is important because many modelling techniques such as ANFIS are dimensional dependent (Choubin et al., 2018; Dubois and Prade, 1980; Yildirim and Bayrujamoglu, 2006). Many model parameters can decrease the model transparency (MacLeod et al., 2010). Incorporating only the important variables provides a simpler, more useful, and more reliable predictive model with enhanced performance and generalization capability (Saeys et al., 2007). Theoretical, suppose dataset $\{ Q = (x_1, y_1, \ldots, x_\mathrm{i}, y_\mathrm{i},)\}$ belong to n-dimensional space, the FS process transforms the dataset Q into a new dataset $\{ R = (x_1, y_1, \ldots, x_\mathrm{i}, y_\mathrm{i},)\}$ with dimension $\{J \mid J << \mathrm{n}\}$ that constitutes only of predominant variables and the same number of entries (Bolón-Canedo et al., 2013; Guyon and Elisseeff, 2003; Saeys et al., 2007).

Finding optimal variables for building a model can be problematic and this process can be achieved using several approaches. Widely used techniques for the environmental domain include grey relation coefficients (GRC) (Wang et al., 2013), partial rank

correlation coefficient (PRCC) (Khoshroo et al., 2018), and multilinear regression (MLR) (Chen et al., 2019; Lee et al., 2016), Akaike's information criterion (AIC) (Akaike, 1974), random forestry feature importance (RFFI) (Ban et al., 2018; Kerckhoffs et al., 2019). However, Akaike's information criterion (AIC), MLR, Spearman, Pearson, and PRCC have limited application confined to variables that have underlying linear relationships (Alimissis et al., 2018) – a phenomenon uncommon for variables that influence the transformation of ENPs in aqueous media. Also, they are computationally prohibitive for problems involving highly multidimensional variables (Akaike 1974; Chiu, 1996). Therefore, the FS was investigated using a combination of MLR, gamma statistic test (GST) (Stefánsson et al., 1997), permutation accuracy importance measurement (PAIM) (Matin et al., 2018) and XGB feature importance. Multiple techniques were used to ensure essential information useful in decision-making was not lost during this process (Chang et al., 2014).

3.4.1.1.5.1 Gamma test (GT)

GT is a non-linear modelling data analytic technique used to evaluate the sensitivity of input series (Evans and Jones, 2002; Stefánsson et al., 1997). It is an estimate of the model output's variance accounted through an unknown smooth non-linear function (Noori et al., 2011). Here, only salient aspects of GT are highlighted, and details are well articulated elsewhere (Evans and Jones, 2002; Stefánsson et al., 1997).

Suppose that the generic relationship of the input(s) and output of M data points is defined as:

$$y = f(\mathrm{x_i} \ldots \mathrm{x_d}) + r \tag{3.5}$$

where f is an unknown smooth function that maps inputs vector-x to output y, and r represents the noise with a mean distribution of zero. Thus, GT can be determined using k^{th} nearest neighbor using the expressions:

$$\delta_M(k) = \frac{1}{\mathrm{M}} \sum_{k=0}^{M} |\mathrm{x_{N(i,k)}} - \mathrm{x_i}|^2 \ (1 \leq k \leq p \tag{3.6}$$

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=0}^{M} |y_{N(i,k)} - y_i|^2 \quad (1 \leq k \leq p) \tag{3.7}$$

with $N_{(i,k)}$ representing the index of the k-th closest neighbor to the x value, $y_{N\,(i,\,k)}$ denoting the corresponding output, |…| as the Euclidean distance, M being the length of the data, p predetermined as equivalent to 10 (Stefánsson et al., 1997), and $\delta_M(k)$ signifying the average square distance to the k^{th} nearest neighbor. The linear regression of the Gamma statistic (Γ) is defined using the expression:

$$\gamma = B\delta + \Gamma \tag{3.8}$$

With B as the gradient. If the x-intercept δ = 0 then the gamma value is equal to the error variance. Theoretically, for a given number (*n*) of variables, there are generally (2^n–1) input plausible combinations. The GT was applied using winGamma software version 1.97.

Algorithm 3 Gamma test

1. Upload dataset { $Q = (x_1, y_1, \ldots, x_i, y_i,)$} into gamma software
2. Set |…| as the Euclidean distance (10)
3. Compute the Γ, SE, V-ratio
   3.1. Using a complete set of inputs (n),
   3.2. Eliminate one input at each cycle (n-1)
   .
   .
   .
   3.3. Repeat until all inputs have been removed
4. Compare the gamma static values

The input variables were selected through an elimination approach excellently discussed **Algorithm 3** and applied in previous works (Moghaddamnia et al., 2009; Remesan and Mathew, 2015). The Γ provides a measure of the best achievable MSE, V-ratio signifies a degree of predictability, and the standard error (SE) illustrates the reliability of the Γ value (Moghaddamnia et al., 2009; Remesan and Mathew, 2015). A Euclidean distance of 10 was assigned. To rank the variable's importance, the

procedure applied was as follows: the Gamma value was computed using the complete set of inputs. Then, each variable was omitted and subsequently, the gamma values were computed using the remaining variables in this case (n-1). This procedure was repeated until each of the variables was eliminated (step 3c). Subsequently, the resulting variables were compared with the gamma values of a complete set. If the Γ values from the omitted value were less than the Γ value of the complete set ($\Gamma_{n-1} < \Gamma_n$) then the omitted had low importance, otherwise ($\Gamma_{n-1} > \Gamma_n$), the contrary holds the variables were included in subset R.

3.4.1.1.5.2 Permutation accuracy importance measurement

PAIM is an inbuilt system in an RF and is used to rank candidate predictors in high-dimensional data (Matin et al., 2018). PAIM can reduce overfitting and depict univariate and multivariate interactions compared to the mean decrease impurity mechanism (Janitza et al., 2016). Herein, only salient features of PAIM are highlighted, and details are well articulated elsewhere (Matin et al., 2018). The permutation importance of the variable x_j is defined as:

$$VI_j^{\pi} = \sum_{i=1}^{N} \mathrm{L}(y_i, f(x_i^{\pi,j}) - \mathrm{L}(y_i, f(x_i)) \quad (3.9)$$

Where the $x^{\pi,j}$ represents the matrix achieved by randomly permuting the j^{th} column and $\mathrm{L}(y_i, f(x_i))$ is the loss function between predicting y_i and $f(x_i)$ of the j^{th} feature (Strobl et al., 2008, 2007).

Algorithm 4 Permutation accuracy importance measurement

1. Load the dataset { $Q = (x_1, y_1, \ldots, x_i, y_i,)$}
2. Initialize the model
3. Train the model on the original dataset
4. Evaluate the accuracy of the original dataset
5. Initialize an array to store feature importance scores
6: Loop through each feature
    6.1. Permute the values of the selected feature
    6.1. Evaluate the accuracy of the permuted dataset
    6.1. Calculate the feature importance score
7. Display or use the feature importance scores

To determine the importance of a variable x_j, the approach uses out-of-bag samples based on trees that are not trained (x_i, y_i) (Janitza et al., 2016). PAIM was implemented using the R software package following the procedure described by Matin et al. (2018) and summarised in **Algorithm 4**

3.4.1.1.6 Data splitting and class balancing

The size of the training data set has a significant impact on the prediction capabilities of the developed ML algorithm. Splitting data into the training dataset and to an independent test subset is a critical step for unbiased data analysis (Stafoggia et al., 2019). A common challenge in data splitting is dealing with unbalanced data and selecting a suitable split ratio (Subramanian and Palaniappan, 2021). Unbalanced data occur when either training or testing data has unequal representations in either of the data sets. There are two important aspects to consider when deciding on suitable split ratios: parameter estimates result in greater variance with fewer training data. Whereas, the performance statistic results in greater variance with fewer testing data. So, the data should be divided in such a way that neither occurs (Balraadjsing et al., 2022; Furxhi et al., 2019a).

Pareto principle (80:20) is the widely applied ratio (Takahashi and Takahashi, 2019), however, other ratios, i.e. 90:10 (Findlay et al., 2018), 60:40 (Balraadjsing et al., 2022) and 70:30 (Papa et al., 2015) have been explored. The optimal split ratio is influenced by the size of the data set considered. For "smaller" datasets (n < 1000) the split ratios 70:30 and 60:40 are usually sufficient to create independent data sets that have an adequate representation so that neither the parameter estimates nor the performance statistic results in greater variance. Various statistical sampling techniques (e.g., stratified sampling, simple random sampling, etc.) exist to partition the data into training and testing sets (Reitermanova, 2010). 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). The larger fraction (70%) was used to train the models and the ‘out of bag’ samples (30%) for validation purposes.

3.4.1.2 ML training process

Six different ML algorithms, namely MLR, ANFIS, SVR, XGBoost, ANN, and RFR were supervised to predict the transformation processes of ENPs in the aqueous environment. The SVR, MLR, XGBoost, and RFR algorithms were implemented in Scikit-Learn Library in Python v3.0. The implementation of the ML algorithm has been simplified through the use of the Scikit-Learn library for Python (Pedregosa et al., 2011). ANN algorithm was built using the TensorFlow library in Python v3.0. ANFIS algorithm was trained in the ANFIS GIU toolbox in MATLAB R2007, software. Both Violin plots and Taylor diagrams were applied using R-studio software. Various optimizers were applied such that the predicted value f(x_i) was close to y_i, so that error converges to zero.

3.4.1.2.1 Artificial neural network

Feed-forward NN includes multilayer perceptron (MLP), single-layer perceptron (SLP), and radial basis function neural networks (RBFNN). Figure 3.4 shows the structure of the MLP. It is a three-layer structure with hidden (h), outputs coupled by neurons, and input variables (Nielsen, 2015; Theodoridis, 2015). The input variables that feed the NN make up the first layer. The second layer or hidden layer (s) acts as the connector and minimises the overlay between patterns. The strength of connection to neurons is determined by the product of weights (w_1,…,w_n) and inputs ($x_1, \ldots, x_n$) (Nagy, 1991; Sahoo et al., 2006; Van Der Malsburg, 1986). The output of the MLP is defined by the function expressed in Equation 3.10 (Dogan et al., 2009; Nielsen, 2015).

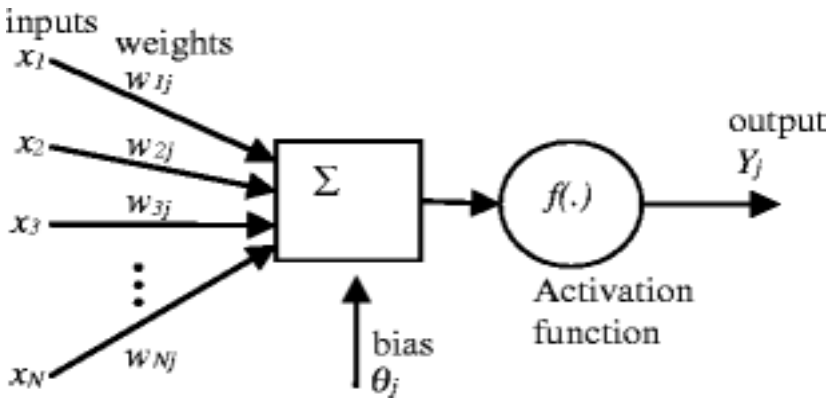


Figure 3. 4. Basic structure of MLP with inputs, weights, hidden layer, and an output connected by neurons. Adapted from (Ahmed et al., 2013). Copyright © 2013, with permission from Springer Nature.

$$f(x)=\sum_{j=1}^{J}\omega_j\delta\left(\sum_{i=1}^{I}w_{ij}\delta(x_i)+\alpha_j\right)+\beta+\varepsilon \tag{3.10}$$

Where $x=(x_1,x_2,\ldots,x_n)$ is a k-dimensional vector of the input unit that represents the raw information that is fed into the network, $\delta(x_i)$ represents activation functions (AFs) that transform the output signals from individual neurons to fit within the active regions, w_{ij} is the weight factor on the connection between the Ith input neuron and jth hidden neuron, α_j is the bias in the jth hidden neuron, w_j is the weight on the connection between the jth hidden neuron and the output neuron, β is the bias of the output neuron, ε is a random error with a mean of zero. I and J are the number of inputs and hidden neurons (Ahmad et al., 2013).

Learning in ANN is influenced by numerous aspects such as input-output parameters, learning rate, activation/transfer functions, the number of neurons, epochs, etc. Identifying, the appropriate hyperparameters requires patience and "try and error" (Rehman et al., 2016; Shabanzadeh et al., 2015). The hyper-parameters differ depending on the complexity of the problem. In this work, key steps followed for developing MLP are summarised using pseudocode in **Algorithm 5**. The ANN was trained using feed-forward (FF) and back-propagation (BB) mechanisms. In this approach, the MLP-ANN was subjected to a hybrid learning algorithm where least squares approximation (LSE) and gradient descent were implemented.

Algorithm 5 Multilayer perceptron – artificial neuron network

Parameters:

- Number of hidden layers (N_{layer})
- Number of neurons in the hidden layer ($N_{layerneurons}$)
- Weight (*w*) and bias (*b*)
- Number of training epochs (epochs)

Procedure:

1. Load the training dataset (X, y)
2. Designing the ANN architecture
   2.1 Assign the input, and output neuron (s)
   2.2 Assign the N_{layer}

3. Feedforward mechanism:
   3.1. Initialize w and b
   3.2. Compute the weighted sum
   3.3. Transform the output signals from individual neurons using various AFs
      − $y = \frac{1}{1+e^{-x}},\ y \in [0,1]$ sigmoid
      − $y = \frac{1-e^{-2x}}{1+e^{-2x}}, y \in [-1,1]$ hyperbolic tangent
      − $y = 0,\ for\ x < 0; x \geq 0,\ f(x) \in [1,\infty]$ RELU
   3.4. Compute the error:
4. Backpropagation mechanism:
   4.1 Calculate the gradient of the loss for the weights and biases.
   4.2 Update weights and biases using gradient descent:
5. Repeat steps 2 and 3 for each:
   − epoch (from 1 to epochs)
   − determine (2 from to $N_{neurons}$)
6. Save the trained ANN models

The random weights and biases were generated in the FF mechanism. The number of neurons in ANN was equivalent to the dimension of the dataset. The number of neurons was equal to the number of features in the input layer and equal to one in the output layer since all developed models were regression problems. A single hidden layer was used for the ANN structure. The number of epochs (0, 10, 100, 500, 1000) and neurons $(z_{i1}, z_{i2}, \ldots, z_{ik})$ in the hidden layer (maximum given by 2^N-1, N the number of inputs) were investigated until the least measured error was obtained for the suitable structure. To transform the output signals from individual neurons to fit within the active regions both sigmoid-based (e.g. hyperbolic tangent (tanh) and the logistic sigmoid (logsig) and non-sigmoid (e.g. rectified linear unit (ReLU) activation functions (AFs) (He et al., 2015; Krizhevsky et al., 2012; Maas et al., 2013) were compared in the hidden layer. The pure-linear (purelin) previously successful in numerous studies (Alade et al., 2020; Vieira et al., 2018) was applied in the output layer.

$$w^* = argmin_w \sum_{n=1}^{N} loss\left(L^{(n)}, y^{(n)}\right) \tag{3.11}$$

$$Squared\ loss: \sum_{i}^{N} \frac{1}{2}\left(L_i{}^{(n)}, y_i{}^{(n)}\right)^2 \quad (3.12)$$

Where L = f(x,w) is the output of the neural network, L is the output of the feedforward with random weights and the output of training data.

$$w^{y+1}=w^{y} - \eta \frac{\partial E}{\partial w^{y}} \quad (3.13)$$

$$b^{y+1}=b^{y} - \eta \frac{\partial E}{\partial b^{y}} \quad (3.14)$$

Where η the learning rate and *w* represents the weight and E is error/loss, then y is the output value

The difference between the expected and desired outputs is known as the loss or cost function. The error/loss function between the target and predicted values were calculated after a successful optimisation of the weights (randomised values) for every connection (Srivastava et al., 2014). Subsequently, a backward pass was done where the parameters were updated using gradient descent as an adaptive approach (Zara et al., 2019). The loss function and LSE are expressed in Equations 3.11 and 3.12. Several backpropagation approaches have been reported to determine the cost function. These include the conjugate gradient algorithm (Burney et al., 2007), the Quasi-newton method (Robitaille et al., 1970), Levenberg-Marquart (Lourakis, 2005), and Gradient descent (GD) (Burney et al., 2007; Men et al., 2007). The GD as the first derivative can be expressed as shown in Equations 3.13 and 3.14. Both adaptive momentum (Adam) and stochastic gradient descent (GD) were investigated as optimisers. The learning rate (η), and the momentum (α) were set as 0.1 and 0.09, respectively to reduce overfitting for the gradient descent method (Alade et al., 2020).

3.4.1.2.2 Support vector regression

SVR is a component of SVM (Schölkopf and Smola, 2002; Zarei et al., 2018). SVR relies on the idea of statistical learning, which applies both linear and non-linear regression to solve problems in quadratic multidimensional feature space by use of a hyperplane (Valizadeh and Sohrabi, 2018; Vapnik, 1995). The linear function f(x) of the SVR is expressed as;

$$f(x) = (\omega, x_i) + b \tag{3.15}$$

Where (ω, x_i) represents the dot product among vector ω and x_i, $\omega \in R^k$ represent the weight vector, $b \in R$ the threshold (Rui et al., 2019; Valizadeh and Sohrabi, 2018).

The Euclidean norm is represented by the regulation factor (C) and term ω^2 in convex optimization, which integrates non-linear surfaces. It can be explained as follows:

$$min\frac{1}{2}\|\omega^2\| + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{3.16}$$

$$subject\ to \begin{cases} y - \omega x_i - b \leq \varepsilon + \xi_i \\ \omega x_i + b - y \leq \varepsilon + \xi_i^* \\ \xi_i \xi_i^* \geq 0, i = 1, \ldots, n \end{cases}$$

Where the $\xi_i \xi_i^*$ are two slack variables.

Equation 3.16 allows objects to reside beyond the sensitivity tube due to the slack variables (ξ, ξ∗∈ n), and the Euclidean norm guarantees that the predicted function is flat. The error is zero if $f(x_i) = y_i$, otherwise, Equation 3.17 can be used to establish the allowable maximum error using epsilon (ε) (Gretton et al., 2012).

$$L(y_i, f(x)) = \begin{cases} 0 & if\ y_i = f(x) \leq \varepsilon \\ |y_i - f(x)| - \varepsilon & otherwise \end{cases} \tag{3.17}$$

For higher dimensional feature space, the function is mapped with Lagrangian dual representation as follows

$$\max_{\alpha,\ \alpha^*} \quad \sum_{i=1}^{n} y_i(\alpha_i + \alpha_i^*) - \varepsilon \sum_{i=1}^{n}(\alpha_i + \alpha_i^*) \tag{3.18}$$

$$-\frac{1}{2}\sum_{s=1}^{n}\sum_{i=1}^{n}(\alpha_i + \alpha_i^*)(\alpha_i + \alpha_i^*)K(x_i, x)$$

$$s.t. \quad \sum_{i=1}^{n}(\alpha_i + \alpha_i^*) = 0,$$

Where the α_i and α_i^* in Equation 3.18 represents the Lagrange multiplies (Smola et al., 1998).

The performance of SVR models is dependent on numerous aspects including the kernel functions, penalty coefficient (C), epsilon (ε), and gamma (γ). The C estimates the degree of tolerance greater than ε and the complexity of the model. The complexity

of the model and the degree of tolerance of more than ε are estimated by the value of C (Chen et al., 2019; Xue et al., 2020). The width of the tube surrounding the estimated function (hyperplane) is determined by the value of ε (Zarei et al., 2018). The γ coefficient determines the radius of influence of a single training point or approximates the curvature of a decision boundary. Unlike the polynomial or linear kernel, this parameter only applies to the Gaussian RBF kernel.

A wide and small similarity radius is indicated by low and high values of γ, respectively. Overfitting is common in models with extremely high γ values (Gretton et al., 2012; Smola and Schölkopf, 2004). The key steps followed for developing SVR are summarised using the pseudocode described in **Algorithm 6**. SVR was built from the standard kernels of the polynomial (poly) and radial basis function (RBF) (Choubin et al., 2018; Hou et al., 2020). Parameters of ε and γ were varied between 0.1–0.3 and 1–10, respectively, to optimise the model performance. The C was set at 1 and a tolerance of 0.001 was employed to train the SVR models until the model with the least bias was obtained.

Algorithm 6 Support vector regression

Parameters:

- gamma (γ)
- penalty coefficient (C)
- epsilon (ε)
- Learning rate (η)
- Max _iter= -1 and tol=0.001
- E is the error
- y is the actual output value
- L = f(x) is the predicted output

Procedure:

1. Load the training dataset (X, y)
2. Initialize dual coefficients γ and error terms (ε)
3. Calculate the predicted output using different kernel functions $\mathrm{k}(x_i, x)$
   - $(\alpha {x_i}^T + \alpha_i)^d$ polynomial function
   - $-exp\left\{-\frac{|x_i+x|^2}{2\sigma^2}\right\}$ radial basis function
4. Compute the error (residual):
   - Error = y_i - L

5. Update γ and error terms (ε)
6. Repeat 3-5 until all variations of γ and error terms (ε) are made
7. Save the trained RF model

3.4.1.2.3 Random forest regression

RF constitutes leaf, internal, and root nodes. The use of DTs allows the RF-based models to integrate and/or handle data defined by uncertainties, and ambiguities. RF can be applied to the classification and regression of real-life ecological problems (Mirzaei et al., 2021; Trinh et al., 2018). Predictions are made by aggregating the predictions of n_{tree} (e.g., a majority vote for classification and an average for regression) (Cutler et al., 2012). The loss square error (LSE) is expressed in binary form using the Boolean system of 0 or 1, as described in Equation 3.19.

$$L\big(y, f(x)\big) = \begin{cases} 0 \; if \; y = f(x) \\ 1 \; otherwise \end{cases} \tag{3.19}$$

The outputs of RF regression and classification are determined by averaging the outputs of each DT (Equation 3.20) and using Gini impurity (Equation 3.21), respectively.

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} h_{\mathrm{J}}(x) \tag{3.20}$$

$$f(x) = arg\max_{y\in\omega} \sum_{j=1}^{J} \mathrm{I}\,(\,\mathrm{y} = h_{\mathrm{J}}(x)) \tag{3.21}$$

where $h_1(x), \ldots, h_{\mathrm{j}}(x)$ denote leaf nodes or the base nodes (trees) (Biau, 2012; Biau and Devroye, 2010). For minimizing the LSE for regression function and classification the conditions are defined in Equations 3.22 and 3.23 (referred to as the Bayes rule), respectively.

$$f(x) = E(Y|X = x) \tag{3.22}$$

$$f(x) = arg\max_{y\in\omega} P\,(Y = y|X = x) \tag{3.23}$$

Where Y is denoted by ω

The key steps followed for developing RFR are summarised using pseudocode described in **Algorithm 7**. The RF has major two parameters; n_{tree} (the number of trees used in the forest) and m_{try} (the number of random variables used in each tree). The samples were bootstrapped from the original data series. Bagging is applied to introduce the randomization and growth of each decision tree (bootstrapped sample) (Ban et al., 2018; Liaw and Wiener, 2002). In this study, the number of decision trees investigated ranged from 0 to 500 using a randomised default state of 42. Splitting was halted when a leaf node/terminal node achieved the desired accuracy. Prediction output was determined by averaging DTs (Liaw and wiener, 2002).

Algorithm 7 Random forest regression.

Inputs:

- Number of trees (n_{tree})
- Number of features to consider at each split (m_{try})
- Maximum depth of each tree (max_{depth})

Procedure:

1. Load the training dataset (X, y)
2. Create a bootstrap sample from the original data series
3. Create a decision tree using the bootstrapped sample
    3.1. Split nodes based on a randomly selected subset of features (m_{try}).
    3.2. Grow the tree until a maximum depth is reached (max_{depth}).
4. Repeat for 1 and 2 until (from 1 to n_{tree}) are added in the forest
5. The final prediction was the average for regression.
6. Save the trained RF model

3.4.1.2.4 Adaptive neuro-fuzzy inference systems

ANFIS consists of five phases (Figure 3.5) and a detailed discussion on the individual layers see (Gong et al., 2016; Jang et al., 1997; Zaghloul et al., 2020). Only important aspects are included here for ease of reading. **Layer 1**: Consists of adaptive nodes from input x and y, with MF(μ) using linguistic labels A and B, respectively. The number of nodes in layer 1 is expressed as the product of the number of inputs and MFs for each predictor variable (Buragohain and Mahanta, 2008). **Layer 2***:* The rule's strength is provided by two fixed nodes labelled with Π, which multiply the resultant value by

the min (AND) operator. **Layer 3**: Fixed node labels (N) that normalize firing strength. ***Layer 4***: Adaptive nodes, which are linear functions with function coefficients modified by the ML feedforward network's error function. ***Layer 5***: Depicts the sum of net outputs of all incoming signals of the nodes using the Sugeno fuzzy inference rules (Jang, 1993).

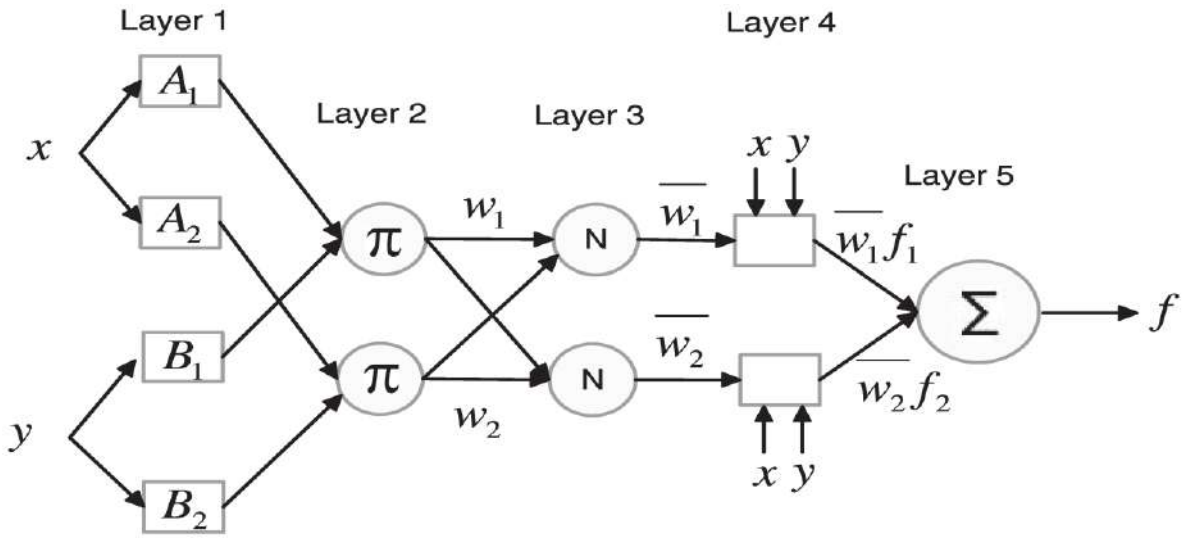


Figure 3. 5. General architecture for ANFIS. Adapted with permission from (Loukas, 2001). Copyright © 2021, American Chemical Society.

In Takagi–Sugeno ANFIS the outputs expressed in Equation 3.24 can be either linear "first-order Sugeno FIS" or constant numerical "zero-order Sugeno FIS". The Sugeno-type FIS is appropriate with a linear function and traditional quantitative-based domain because it uses first-order equations (Jang et al., 1997; Takagi and Sugeno, 1985; Taniguchi et al., 2001).

$$\text{if x is A and y is B then } f_1 = px + qy + r \tag{3.24}$$

Input x and y and A and B are MFs. where p, q, and r are the consequent parameters of the linear output functions (f_i) (Taniguchi et al., 2001).

The ANFIS depends on several hyper-parameters namely; the input and output fuzzy sets, the number of fuzzy rules, data clustering methods, epochs, error rate, etc. A pseudocode to summarise the key steps for developing ANFIS is described in **Algorithm 8**. The training and testing datasets were loaded into ANFIS GUI in MATLAB. Several data clustering techniques have been reported, such as grid

partitioning (GP) (Jang, 1993), mountain methods (Yager and Filev, 1994), and subtractive clustering (SC) (Chiu, 1994). The GP and SC are the most commonly used approaches built into the ANFIS editor in MATLAB. The expression below can be used to predict the number of rules for grid partitioning based on the number of input variables and membership functions.

$$N_{Rules} = M_{MF}{}^{n_{inputs}} \tag{3.25}$$

The GP is suitable for less than 5. The number of fuzzy rules increases exponentially depending on the number of inputs and the number of MFs as described by Equation 3.25. On the other side, the SC is mostly preferred when the number of input parameters is more than 6 (Kayadelen et al., 2009). The GP approach was chosen herein as it allows the variation of membership functions (MFs) such as triangular, trapezoidal, generalized bell, and Gaussian I and II. The different MFs were assigned to adaptive nodes of inputs $\{ R = (x_1, y_1, \ldots, x_i, y_i,)\}$ with MF (μ) using linguistic labels of A and B, respectively. The parameters of these MFs are termed premises (Buragohain and Mahanta, 2008).

Algorithm 8 ANFIS based on gradient descent and least-squares methods

Parameters:

- Learning rate (η)
- weight (*w*) and bias (*b*)
- Number of training epochs (epochs)
- Number of MFs (epochs)

Procedure:

1. Upload training dataset (X, y) in the ANFIS GIU
2. Initialize fuzzy sets and consequent parameters
   2.1 Mf type and number of MFs
   2.2 parameters (weights) of the linear combination of fuzzy rules
3. Feedforward pass:
   3.1 Evaluate the MF values for each input
   3.2 Compute the firing strength of each rule
   3.3 Calculate the output of each rule
   3.4 Compute the error between predicted and actual output.
4. Backward pass

4.1 calculate the parameters of the consequent using GD

4.2 Update the weights based on the error and firing strength

$$\overline{w_i} f_i = \overline{w_i}(p_i x + q_i y + r_i)$$

5. Overall output of the model

$$\sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}$$

5. Repeat 2-4 for each

−epoch (from 1 to epochs)

−number of MFs (from 3 to N_{mfs})

−various MFs triangular, trapezoidal, Generalised bell, Gaussian I, and Gaussian

8. Save the trained ANFIS model

The learning of ANFIS models was based on a hybrid algorithm – a combination of the gradient descent and least-squares methods (Choubin et al., 2018; Noori et al., 2011; Pham et al., 2019), similar to the training of ANN. In the forward pass, the premise parameters were kept constant, while the consequent parameters were determined by LSE. Then backward pass was done were the consequent parameters were kept constant and the premise parameters updated using gradient descent. The algorithm was updated until the number of epochs with the minimum error was achieved. An epoch is a cycle from feedforward to backward pass. Hyper-parameters such as the number of MFs and epochs were investigated in the range of (3-5) and (50, 100, and 600), respectively, based on a tolerance of 0.001.

3.4.1.2.5 Multiple linear regression (MLR)

Multiple regression is the extension of simple linear regression (LR) used to (i) predict model output(s), and (ii) determine the most significant variable among the predictors toward output (Choubin et al., 2018; Lee et al., 2016; Rajaee et al., 2009). MLP with two explanatory variables (X_1, X_2) and the response variable (Y) can be illustrated as follows.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \epsilon \tag{3.26}$$

Where ϵ the residual term of the model and βi is the regression coefficients. A general procedure to estimate the regression coefficients is minimisation using the expression.

$$\sum_{i=1}^{n} M(\epsilon_i) = \sum_{i=1}^{n} M(Y_i - \beta_0 - \beta_1 \cdot X_{1,i} - \beta_2 \cdot X_{2,i}) \tag{3.27}$$

Where Y_i, $X_{1,i}$, $X_{2,i}$, and $\epsilon_i$denote the i^{th} residual being minimized through the following, $M(x) = x^2$, which is also known as estimation using the least-squares method. **Algorithm 9** summarises the key steps followed for developing MLR

Algorithm 9 Multiple Linear Regression Training

Parameters:

- E is the error
- y is the actual output value
- L = f(x) is the predicted output

Procedure:

1. Initialize coefficients and intercept:
   - Set initial values for coefficients $\beta_0 + \beta_1 + \beta_2$ and intercept (ϵ).
2. Calculate predictions:
   - Calculate the predicted output using the current coefficients and intercept:

     $\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \epsilon$

   b. Compute the error:
   - Error = y_i - L
3. Repeat for each iteration until convergence:
4. Save the trained MLR model

3.4.1.2.6 Extreme Gradient Boosting

XGBoost 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, $\in \mathbb{R}^{n \times m}$ and $y \in \mathbb{R}^{m \times 1}$)}. The mathematical function of XGBoost is expressed as follows:

$$y_j^{(\mathrm{dt})} = \sum_{k=1}^{K} f_k(x_j) \tag{3.28}$$

Where f_k is the independent tree, $F = (f_1, f_2, .., f_k)$ denotes the regression trees, $y_j^{(\mathrm{dt})}$ 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:

$$L^{(\mathrm{dt})} = \sum_{j=1}^{m} l(y_j, y_j^{(\mathrm{dt})}) + \Omega(f_k) \tag{3.29}$$

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

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;

$$L^{(\mathrm{dt})} = \sum_{j=1}^{m} l[(y_j, y_j^{(dt-1)}) + f_{\mathrm{dt}}(x_j)] + \Omega(f_{dt}) \tag{3.30}$$

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

$$L^{(\mathrm{dt})} = \sum_{j=1}^{m} l[\, g_j f_{\mathrm{dt}}(x_j) + \frac{1}{2} h_j f_{dt}^{\,2}(x_j)] + \Omega(f_{dt}) \tag{3.31}$$

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

The scoring function in Equation (3.32) evaluates the optimal weight value to compute the predicted value for each leaf node.

$$L^{(f)} = -\frac{1}{2}\sum_{n=1}^{T} \frac{G_j^{\,2}}{H_j + \lambda} + \gamma T \tag{3.32}$$

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

The accuracy of XGBoost models is rooted in the optimisation of a large variety of hyper-parameters such as learning rate, maximum depth, regularization, or penalty term on weights among others. A pseudocode that summarizes the key steps for developing XGBoost is described in **Algorithm 10.** The boosting approach fits a series of small decision trees in a sequential process, with each tree built after the previous one. Each decision tree is adjusted to the residuals from the model and is added to the fitted function to update the residuals. Trees can be rather small, with just a few terminal nodes, which improves the model (James et al., 2017). The number of n_estimators was investigated from 50, 100, 200, 500, and 1000. Theoretically, a higher number of estimators leads to better performance and reduces the impact of overfitting as the result of increased diversity and robustness. This is generally based on the concept of ensemble learning similar to RF.

Algorithm 10 Extreme Gradient Boosting Training

Parameters:

- Number of tree (n_estimators)
- Learning rate (eta)
- Maximum tree depth (max_depth)
- E is the error
- y is the actual output value
- L = f(x) is the predicted output

Procedure:

1. Load the training dataset (X, y)
2. Initialize a boosted ensemble model with a simple model:
3. Compute the loss function or residual
    - Error = y_i - L
4. Compute the negative gradient of the loss function
5. Fit a weak learner to the negative gradient using the training dataset:
   3.1 Specify parameters such as max_depth, etc.
   3.2 Use the negative gradient as the target variable for the new tree.
6. Update the ensemble model:
   4.1 Compute the weight of the new tree using the (eta).
   4.2 Add the weighted new tree to the ensemble.
6. Repeat for each boosting round (from 1 to n_estimators):
7. Save the trained XGBoost model

The learning rate is used to prevent overfitting by making the boosting process more conservative. The learning rate was investigated between 0.001, 0.01, 0.05, 0.10, 0.20, and 0.30. Moreover, the maximum depth or size of a tree is the number of splits in each tree. Maximum depth controls the complexity of the boosted structure. It is used to control overfitting because higher depth allows the model to learn relationships that are highly specific to a particular sample. The max depth was investigated between 3 -10. Moreover, the regularization term controls the complexity of the model, helping to avoid overfitting (Chen & Benesty, 2016). XGBoost regularizes the weights of variables or, equivalently, shrinks the weights toward zero. Regularization advantage is rooted in the bias-variance trade-off of 1. Controlling the best combination of parameters is necessary to optimise and improve the model and reduce model complexity (Chen et al., 2015). The final model is a linear combination of hundreds or even thousands of trees.

3.4.1.3 Performance assessment criteria

The best-performing ML approach (es) was selected using the k-fold cross-validation (CV) which partitioned the calibration and validation data sets into several folders. The K-folder CV approach promotes a low-bias prediction under limited and insufficient data and reduces over-representation (Alade et al., 2020; Lim et al., 2019). Figure 3. 6 shows the different metric systems for risk of overfitting, nonlinearity, and bias-variance balance (Li et al., 2022). Many research studies use k = 5 or 10 (Findlay et al., 2018; Subramanian and Palaniappan, 2021). The larger, the value of k the better the results of the trained model and the low bias and variance (Balraadjsing et al., 2022; Ban et al., 2020; Choi et al., 2018). In this study, the data set was arbitrarily divided into ten equivalent sub-samples, nine for training purposes, and one for validation.

Figure 3. 6. Diagram elucidating risk of overfitting, nonlinearity, and bias-variance balance. Adapted from (Li et al., 2022). Copyright 2022, with permission from Elsevier.

Several statistical metrics criteria such as root mean squared error (RMSE), mean absolute error (MAE), Nash Sutcliffe efficiency rating (NSE), correlation coefficient (R), and coefficient of determinants (R^2) in Equations **3.33-3.37,** respectively were utilised to determine the Goodness-of-fit. Different metric parameters assumed variant values. RSME determines the square root of the residual between the real value and the predicted model output. Smaller values of MAE and RMSE signify the better performance of the model, whereas high values indicate outliers in the data set (Alexander et al., 2015; Moriasi et al., 2007; Hou et al., 2020; Kozleski, 2017).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i}}^{\mathrm{n}}(\mathrm{t_i} - y_i)^2}{\mathrm{n}}} \tag{3.33}$$

$$\mathrm{MAE} = \frac{\sum_{\mathrm{i}}^{\mathrm{n}}|(\mathrm{t_i} - y_i|}{\mathrm{n}} \tag{3.34}$$

The Pearson R values are vector components described in absolute terms in the range (-1 to 1), with an R approaching the absolute value of 1, indicative of linear agreement, not the direction of magnitude (positive or negative) (Furxhi et al., 2019b). The coefficient of determination (R^2) was determined by squaring R values and in practice, the R^2 is lower than R values. Higher values (closer to 1) of R^2 and NSE signify the better performance of the model and are an indication of a strong linear relationship between measured and predicted values.

$$NSE = \left(1 - \frac{\sum_{i=1}^{n}(t_i - y_i)^2}{\sum_{i=1}^{n}(t_i - \mathrm{Z})^2}\right) \quad (3.35)$$

$$\mathrm{R} = \frac{\sum_{\mathrm{i}}^{\mathrm{n}}(\mathrm{t_i} - \mathrm{Z})(y_i - \mathrm{K})}{\sqrt{\sum_{\mathrm{i}}^{\mathrm{n}}(\mathrm{t_i} - \mathrm{Z})^2} \cdot \sqrt{\sum_{\mathrm{i}}^{\mathrm{n}}(y_i - \mathrm{K})^2}} \quad (3.36)$$

$$R^2 = \left(\frac{\sum_{\mathrm{i}}^{\mathrm{n}}(\mathrm{t_i} - \mathrm{Z})(y_i - \mathrm{K})}{\sqrt{\sum_{\mathrm{i}}^{\mathrm{n}}(\mathrm{t_i} - \mathrm{Z})^2} \cdot \sqrt{\sum_{\mathrm{i}}^{\mathrm{n}}(y_i - \mathrm{K})^2}}\right)^2 \quad (3.37)$$

Where K is the mean of the predicted output, n is the number of samples, $\mathrm{y_i}$ is the predicted output, $\mathrm{t_i}$ is the observed output and Z is the mean of the observed output.

Table 3. 1**.** Model performance rating based on NSE and R^2

Model	NSE value	R^2
Very Good	(0.75 -1)	(0.85-1.00)
Good	(0.65-0.75)	(0.70-0.85),
Satisfactory	(0.50-0.65)	(0.60-0.70)
Acceptable	(0.40-0.60),	(0.40-0.60),
Unsatisfactory	NSE ≤ 0.50	(R^2 ≤ 0.40)

(Adopted from Moriasi et al., 2007; Alexander et al., 2015; Pradhan et al., 2020; Rauf et al., 2018).

3.4.1.4 Performance visualization

3.4.1.4.1 Taylor diagram

Taylor diagram (Taylor, 2001) summarises the performance metrics of the developed model. Taylor diagram entails the use of three statistics namely: standard deviation (SD) of predicted and reference field, centered root mean squared (RMSE), and Pearson R. Mathematically, the relationship between the three statistics is related by the expression using Equation 3.38

$$\mathrm{cRMSE}^2 = {\sigma_p}^2 + {\sigma_o}^2 - 2\sigma_p\sigma_o R \quad (3.38)$$

Where cRMSE, σ_p, σ_o and R represents the centered root mean squared error, the standard deviation of the predicted values, the standard deviation of observed values, and the correlation coefficient, respectively (Taylor, 2001). `

The centered root-mean-square (RMSE) is the difference between the simulated and observed patterns and is proportional to the distance to the "observed" point. The standard deviation of the simulated pattern is proportional to the radial distance from the origin. Simulated patterns that agree well with observations will lie nearest the point marked "observed". These models will have relatively high correlation and low RMSE errors. In the Taylor diagram, the best models were classified using the following criteria: (a) the cosine of the azimuthal angle in a polar coordinate R should be close to 1 and the lower centered RMSE values (b) the amplitude of SD of predicted should be within proximate to SD of reference field (observation) (Taylor, 2001).

3.4.1.4.2 Violin plots

Violin plots (Hintze and Nelson, 1998) are a hybrid of box plots (Tukey, 1977) and kernel density (Benjarnini, 1988). They were used for the visualisation of data distribution and provided an insightful comparison of the developed model against the reference dataset (Choubin et al., 2018). Violin plots consist of four main features of data: mean, distribution, asymmetry, and outliers. White dots on violin plots depict the mean of each dataset and wide regime; signifying high probability distribution, whereas the skinner regime; signifies low probability distribution. The Interquartile range (IQR) (25^{th}, 50^{th}, and 75^{th} quartiles). The ends of solid black points depict the highest (95^{th}) and lowest values (5^{th}) (Hintze and Nelson, 1998). The density at point x is defined as the fraction (x/h) of the data values per unit of measurement that fall in an interval and is expressed using Equation 3.39.

$$\mathrm{d(x/h)} = \frac{\sum_{j=1}^{n} \delta_i}{\mathrm{nh}} \tag{3.39}$$

Where n is the sample size, h is the interval width and δ_i is one when the i^{th} data value is in the interval[x - h/2, x + h/2] or zero otherwise.

3.4.1.4.3 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 3.40 and 3.41, 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 (3.40)

H_A : distribution of permuted samples ≠ random sample (3.41)

3.4.1.4.4 Applicable 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.

3.4.2 Knowledge-based systems

Figure 3. 7 depicts the schematic diagram showing the development of KBS. In contrast to ML which requires structured data pairs, KBSs are generally useful if certain aspects of the problem domain are defined including discrete data, non-structured data, qualitative attributes, and limited quantitative data. The link between inputs to output(s) was expressed using heuristics in words (if-then conditional statement) and/or scores using interpretations or expert insight. To develop KBSs, a collection of inputs {X= $(x_1, x_2, \ldots, x_k)$} and output(s) {Y= $(y_1, y_2, \ldots, y_k)$} defined in the corresponding universe of discourses {H=$(h_1, h_2, \ldots, h_k)$} and {M=$(m_1, m_2, \ldots, m_k)$} was considered.

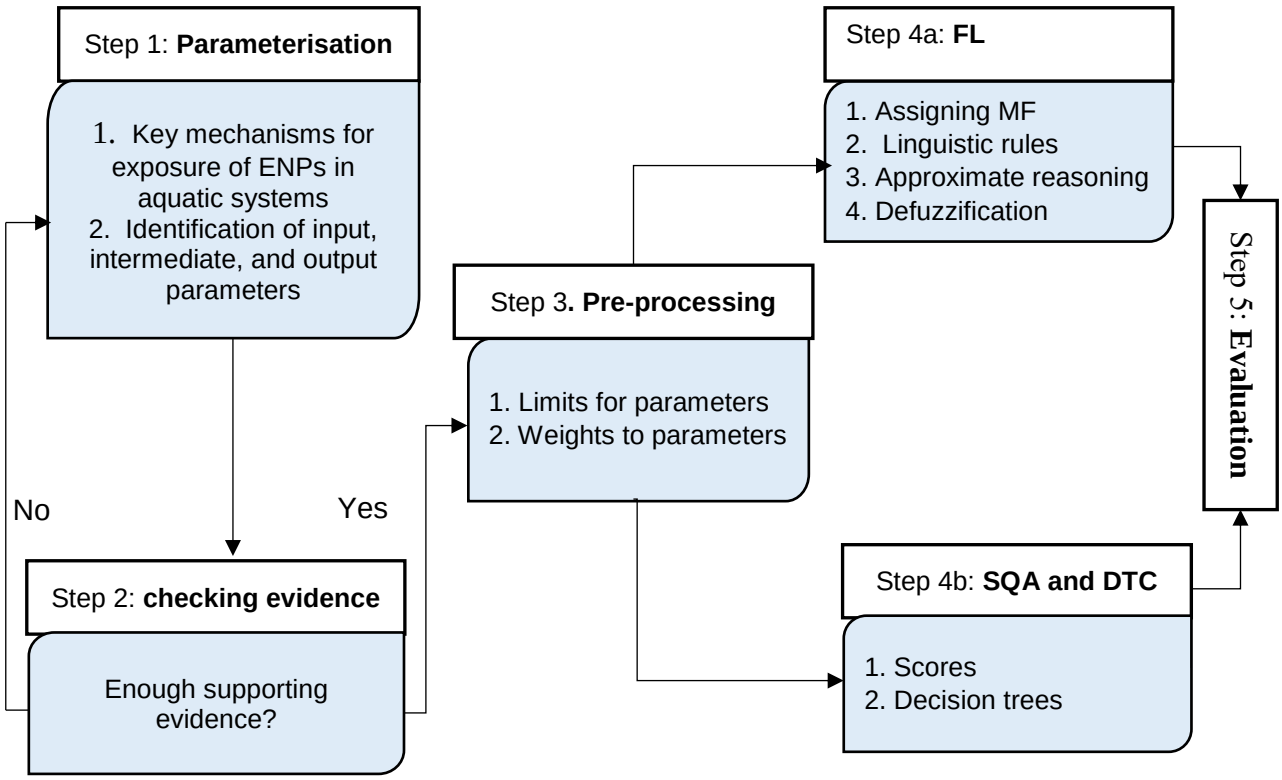


Figure 3. 7: Diagram showing processes followed for the development of KBS (Musee, et al., 2008; Musee, 2017)

3.4.2.1 **Normalisation**

High data variability can complicate model development; therefore, the standardisation process converts the data into a format suitable for modelling purposes. This can help reduce bias, provide consistency, replicate results, and improve transparency (Garca-Diéguez et al., 2015). To render qualitative-based parameters in a format compatible with coding in the fuzzy algorithm, each parameter was assigned numerical indices in the interval 0 to 1 (Musee et al., 2008). The identified inputs were assigned weights and scored using Saaty's scale (Saaty, 1987, 1980). This process used the reviewed insights of the concepts as the basis for the manipulation of scores or weights. These reflected the relative strength of preferences. Higher scores or weights implied a high influence of a given factor on the output under question, and vice-versa holds for lower scores or weights (Saaty, 1987, 1980).

3.4.2.2 Semi-quantitative analysis and Decision Tree Classifiers

The weights assigned were summed as described in Equation 3.42 and Equation 3.43 to develop the leaf nodes of the decision tree classifiers (DTC). The internal nodes, leaf nodes, branches, and root nodes make up the less significant branches DT. The DT begins with a root node. The internal nodes or decision nodes are execution branches between the root and leaf nodes from which the decision is derived (Furxhi et al., 2019; Sizochenko et al., 2019). Using Occam Razor's parsimony principle, a small number of trees can adequately suffice for building a model that can be easily understood and it makes it simpler to generate pure leaf nodes. As a tree becomes larger, it gets more challenging to retain its purity (Huang et al., 2022). Therefore, pruning was used to reduce complexity, and it entailed cutting off branches that were less significant (Dong et al., 2022). A pseudocode to summarise the key steps for developing XGBoost is described in **Algorithm 11.**

$$Max_v = \sum_{n=1}^{K} w_{nH} \tag{3.42}$$

Where w_H is the highest weight assigned to each input, $w = (w1_H, w2_H, .., w_{nH})$ denotes the weights, Max_v is the estimated value after *K-th* weights.

$$Min_v = \sum_{n=1}^{K} w_{nL} \tag{3.43}$$

Where w_L is the lowest weight assigned to each input, $w = (w1_L, w2_L, .., w_{nL})$ denotes the weights, Min_v is the estimated value after *K-th* weight.

$$Median = (Maxv + Minv)\ /2 \tag{3.44}$$

Algorithm 11 Decision trees

```
Function DT ():
    if stopping condition ():
    return Leaf node ()
    best_split = find_best_split()
    left, right = split (best_split)
```

left$_{branch}$ = DT (left)

right$_{branch}$ = DT(right)

return Decision$_{node}$(best$_{split}$, left$_{branch}$, right$_{branch}$)

3.4.2.3 Fuzzy logic model

FISs are classified into three types based on the consequences/output of the fuzzy rules. These are Mamdani-Assilian, Takagi and Sugeno and Tsukamoto FIS. Both Mamdani-type and Takagi–Sugeno are widely applied FISs (Mamdani and Assilian, 1999; Taniguchi et al., 2001). The Mamdani systems are intuitive and best suited for interpretation over Takagi-Sugeno FIS. On the other hand, Sugeno-type FISs are more suitable for dynamic nonlinear systems and have higher computational efficiency. However, they lack the expressive power and human interpretability of Mamdani output (Chaudhari and Patil, 2014). Mamdani-type fuzzy system is shown in Figure 3.8 and has three main components: fuzzification, inference, and defuzzification.

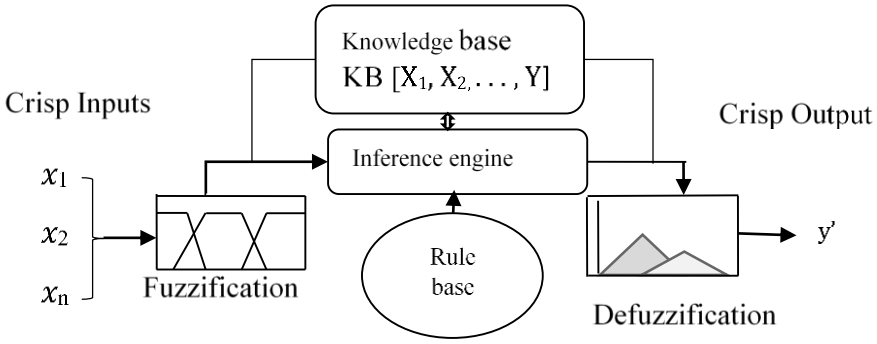


Figure 3. 8. The systematic architecture for Mamdani-Assilian FIS (Musee, et al., 2008).

In this work, the FL was developed using Mamdani and Assilian FIS built-in fuzzy logic toolbox integrated into Matrix Laboratory (MATLAB, R2007) software (Camastra et al., 2015). The key steps followed for developing FL are summarised using the pseudocode described in **Algorithm 12.** The first step involved assigning MFs or LVs and distribution shapes to a collection of inputs(s) {X= $(x_1, x_2, \ldots, x_k)$} and output(s) {Y= $(y_1, y_2, \ldots, y_k)$} defined in the universe of discourses {H=$(h_1, h_2, \ldots, h_k)$} and {M=$(m_1, m_2, \ldots, m_k)$}, respectively. The linguistic variables (LVs) are employed as the

fundamental units of FL (Paul et al., 2018; Zhang et al., 2013; Zhou et al., 2022). According to Lee (1990), a linguistic variable (LV) is a variable whose values are specified in terms of language, or a variable whose value is a fuzzy number.

Generally, there is a paucity of detailed guidelines on the number of LVs or the distribution shapes of MFs; these depend on expert intuition but also on the inference method (Zhou et al., 2022). A high number of sets can yield a high number of rules and increased computation processing (Gacto et al., 2011). Different MFs distribution shapes such as trapezoidal, triangular, Gaussian, sigmoid, piecewise-linear, and bell-shaped MFs have been reported (Fayaz et al., 2017; Johnpaul et al., 2021; Mazhar et al., 2019). The MFs of Gaussian MF, Generalised bell, and Sigmoidal are non-linear and continuous. They have advantages that include being nonzero at all locations, smooth, and succinct (Ali et al., 2015; Ramirez and Mayorga, 2008; Sadollah, 2018). However, because these methods work well for handling statistics and probabilities, they have a longer calculation time (Omid et al., 2010). In this work, the trapezoidal and triangular-shaped MF distributions were carefully chosen based on their high convergence rate, and simplicity of most common varieties of MFs, particularly in real-time applications (Fayaz et al., 2017; Johnpaul et al., 2021; Mazhar et al., 2019).

Algorithm 12 Mamdani fuzzy inference system

Parameters:

- Membership functions (MF)
- Parameters a and d control the size of the trapezoidal base
- Parameters b and c define the "shoulders".
- Parameters a and c control the triangle base and parameter b the peak
- A_{gk} and Bg are the fuzzy set (e.g. low, medium, high)
- Deffuzified output (z*),
- number of rules (n),
- wj is the yield value in the j subset and μ is the MF value of wj

Procedure

1: Open Mamdani FIS GUI system in MATLAB

2. Inputs ($x_k \in H_k$) and outputs ($y_k \in M_k$); H_k and M_k denote the universe of discourse

3. Fuzzification:

- Assign MFs shapes to antecedent and consequent

$\mu(x, a, b, c, d) = \max\left(\min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}, \right) 0\right)$ trapezoidal-shaped MF

$\mu(x,a,b,c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right)$ triangular-shaped MF

3: Linguistic rules for antecedent and consequent

Rg: IF x_1 is A_{g1}. . . AND x_k is A_{gk} THEN y_k is Bg.

4: Aggregation of MFs and rule evaluation method

mini and *max-min* operators

5: Defuzzification techniques based on the COG method

$$z^* = \frac{\sum_{j=1}^{n} \mu_z(w_j)\acute{\omega}_j}{\sum_{j=1}^{n} \mu_z(w_j)}$$

6. Save the trained Mamdani FL model

To link the antecedent and consequent parameters in FIS the syntax of the format 'if' (situation, condition, pattern) and 'then' (actions) rules were used in step three (Camastra et al., 2015; Pepa et al., 2020). The 'if' component of the statement captured the information with different permutations and the 'then' component gave a response. These are linguistic vector inputs MF set to scalar output data in a continuum set of MF (Shi et al., 1999; Traore et al., 2005; Yeung and Tsang, 1997). Theoretically, the maximum number of permissible fuzzy rules in the rule editor interface of FIS are determined based on the Eigenvalue matrix expressed using the exponential function described as $r=w^m$ or $r= w_p...w_n$ where m represents the number of inputs of the system and w the number of the fuzzy sets (Chiu, 1996; Shi et al., 1999; Wu et al., 2000; Yager, 1992).

Moreover, the conclusive results in the FL model from the developed set of fuzzy rules were attained by the inference process known as approximate reasoning. By using this method, it was possible to determine the truth value of one statement based on the truth value of another. Generalized modus ponens (GMP) and generalized modus tollens (GMT) are the two key inference rules (Chaudhari and Patil, 2014; Ramirez and Mayorga, 2008). Several fuzzy rule composition operators are used in the approximate reasoning process, including max-min in Equation 3.45, which represents a low uncertainty range, sum-prod in Equation 3.46, which represents a high uncertainty range. The max-prod is a combination of both max-min and sum-prod (Mamdani, 1974; Mendel, 1995).

The MAX-MIN notation:

$$\text{Fuzzy AND: } \mu_{C1}(z) = \min[\mu_{A1}(x); \mu_{B1}(y)] \quad (3.45)$$
$$\text{Fuzzy OR: } \mu_{C2}(z) = \max[\mu_{A2}(x); \mu_{B2}(y)]$$
$$\text{Fuzzy NOT: } \mu_{C3}(z) = 1-\mu_{A3}(x)$$

The PROD-SUM notation:

$$\text{Fuzzy AND: } \mu_{C1}(z) = \mu_{A1}(x) * \mu_{B1}(y) \quad (3.46)$$
$$\text{Fuzzy OR: } \mu_{C2}(z) = \mu_{A2}(x) * \mu_{B2}(y)$$
$$\text{Fuzzy NOT: } \mu_{C3}(z) = 1-\mu_{A3}(x)$$

Where A_1 and B_1 define the fuzzy sets for both the antecedent and the consequent, respectively. Where $\mu_{C1}(x,y)$, $\mu_{A1}(x)$ and $\mu_{B1}(y)$ are three membership functions of A_1, B_1, and C_1 respectively while *R* and *K* are the universes of a discourse of A1 and B1, respectively

The max-min approach is regarded as the universal approach (Bennajeh et al., 2018; Bouchrika et al., 2014; Pepa et al., 2020). It has advantages such as simplicity, minimizing over-estimation during simulation, and providing the least MF value to limit biases (Mamdani and Assilian, 1999; Mendel, 1995; Zadeh, 1975). The conjugate operator of 'AND' and the *max-min* composite were the preferred methods for approximate reasoning (Pepa et al., 2020). Lastly, the technique of producing crisp output from the FIS following approximate reasoning is known as defuzzification. This procedure is essential to Mamdani FIS and missing from Sugeno-type FIS (Chaudhari and Patil, 2014). The commonly applied types of defuzzification approaches are the mean of maximum (MOM) (Braae and Rutherford, 1978), the height method (HM) (Hellendoorn and Thomas, 1993), and the center of gravity (COG) (Mamdani, 1974) (Yager, 1992, 1992). The mean of the outputs with the highest grades is calculated using the MOM approach (Braae and Rutherford, 1978), while HM is appropriate for the output produced by symmetrical functions (Hellendoorn and Thomas, 1993).

The weighted average of the MF is calculated by the COG and is output as a discrete, crisp arithmetical value (Bockstaller et al., 2017; Kumar et al., 2013; Musee et al., 2008). The output histogram is symmetric and undistorted, and the centroid defuzzification features are consistency, section invariance, monotonicity, and steady and monotonous (Runkler, 1997). The COG defuzzification method is the most applied because it avoids the discontinuities that may appear superior in the MM

defuzzification method (Fayaz et al., 2017; Zhou et al., 2022). The center of gravity (COG) was used as a defuzzification approach (Fayaz et al., 2017; Musee et al., 2008; Zhou et al., 2022).

Chapter 4. Predicting the dynamic aggregation of zinc oxide and titanium dioxide in aqueous systems using machine learning

This chapter describes the results of the application of nonlinear and linear ML (ANN, ANFIS, SVR, RF and MLR) approaches for data mining and prediction of dynamic aggregation of ENPs in freshwater-like systems based on case studies of nZnO and $nTiO_2$

4.1 Introduction

The aggregation is the irreversible clustering of colloid particles to large clusters of agglomerates based on the efficiency of their collision (Dwivedi et al., 2015; Grillo et al., 2015) and is considered the most important process for understanding the behaviour of ENPs. To elucidate the aggregation in aqueous media, theoretical approaches including Derjaguin, Landau, Verwey and Overbeek (DLVO) (Derjaguin, 1941; Verwey and Overbeek, 1955) or (extended DLVO) (Wang et al., 2015) and models integrated with DLVO theory, e.g. Monte Carlo (MC) simulations (Feng et al., 2017), etc. have been reported. Classical DLVO theory, for example, accounts for particle interaction based on van der Waals attraction and electrostatic repulsion forces (Hartmann et al., 2014); whereas the extended DLVO elucidates non-DLVO interactions caused by steric forces (Feng et al., 2017; Wang et al., 2015). However, despite the classical DLVO and linked modelling approaches aiding in elucidating the energy forces that drive the colloidal stability, they are characterised by several limitations. These include only being applicable to describe interactions within a short distance range (0.1–10) nm, spherical shapes, lower concentration of multivalent electrolytes, and, cannot easily elucidate interactions among ENPs and natural colloid particles in multi-component systems (Meesters et al., 2014b; Praetorius et al., 2020; Wang et al., 2015).

This chapter investigated the use of modelling tools such as ML tools to elucidate insights and extract key underlying factors of the ENPs aggregation process using the experimental data generated using various freshwater–like aqueous matrices which are characterised by heterogeneity, multifaceted, uncertainty, multiplicity, sparsity, and noise (Ban et al., 2018; Goldberg et al., 2015b).

4.2 Analysis of data on the dynamic aggregation of ENPs

Data extraction of ENP aggregation in aquatic systems was performed using the process described in Sections 3.1 - 3.3. The search criteria resulted in fifty-four publications discarded and forty-four studies that are summarised in Table A.1. Integrating data from multiple literature studies yielded datasets with 604 and 446 data points for $nTiO_2$ and nZnO, respectively. The datasets had seven continuous features namely; NOM (X_1, mg/L), pH (X_2, dimensionless), IS (X_3, mM)), ZP (X_4, mV), size (X_5, nm)), ENP concentration (X_6, mg/L), and duration (X_7, h)). The hydrodynamic size (D_H) (X_8, nm) was used as the model output. About 98% of ENPs were uncoated, thus coating was completely removed as inputs. Parameters of IS, ZP, pH, NOM, ENP concentration, and time had missing values in the range of 1- 6.29 % as indicated in Table 4.1. Instead of applying the case deletion strategy, the KNN imputation method described in Section 3.4.1.1.3 was used to populate the missing values. The descriptive statistics of datasets are summarized in Table A.2. Figure A.1 and Figure A.2 depict the mass distribution of inputs and output parameters in percentiles.

Table 4. 1. Type of data points and number of missing points for each variable

Variables	Units	Type of data	Percentage of Missing data points (%)	
			$Data_{nZnO}$	$Data_{nTiO2}$
NOM	$mg \cdot \ell^{-1}$	Continuous	1.33	2.00
pH	–	Continuous	-	1.66
IS	mM	Continuous	2.00	3.79
Size	nm	Continuous	-	-
ZP	mV	Continuous	4.26	6.29
ENP concentration	$mg \cdot \ell^{-1}$	Continuous	3.36	--
Time	hour	Continuous	4.48	5.46
Coating	–	Categorical	-	-
HDD	nm	Continuous	-	-

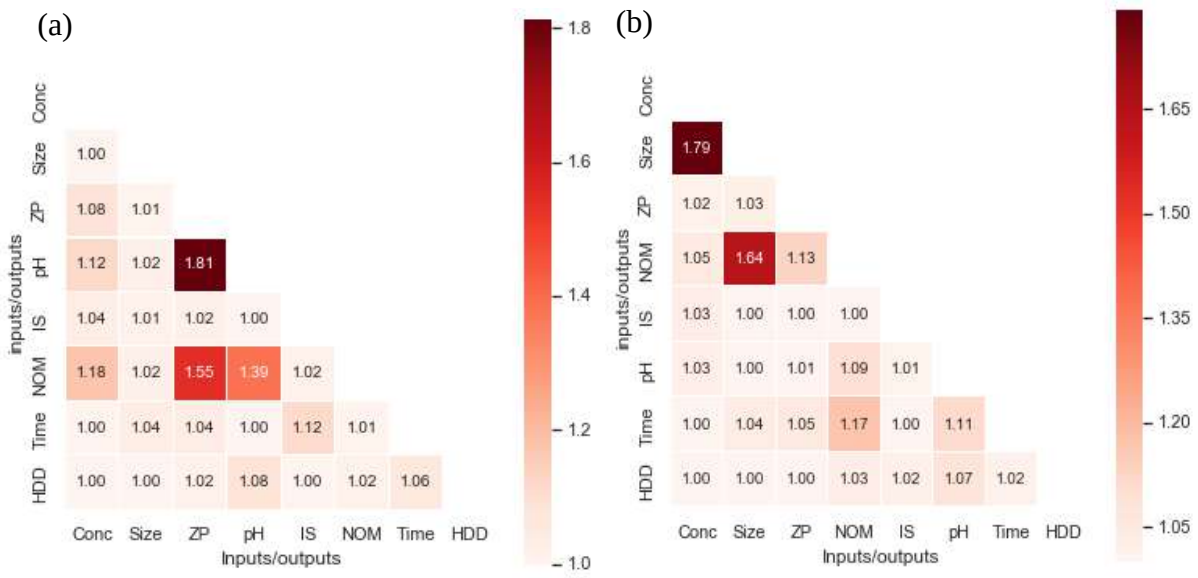


Figure 4. 1. Vif values to estimate the multicollinearity for (a) $nTiO_2$ and (b) nZnO

4.3 Feature selection results

Too few or many model input variables are undesirable. Too many variables can complicate the network model, while a smaller number of variables may not capture the information adequately (Blumer et al., 1987). The results described in Figure 4.1a ($nTiO_2$) and Figure 4.1b (nZnO) showed no multicollinearity among the features as their vif values were < 5. Further, results in Table 4.2 and Figure 4.2 of GT and PAIM, respectively, were used to identify the best combination of the model input parameters. The GT results in Table 4.2 are based on Γ, V-ratio, and SE. The values of Γ_{ZP}, Γ_{pH}, and Γ_{time} were higher than the reference value (Γ_{All}) of 0.063 and 0.076 for $nTiO_2$ and nZnO, respectively.

The results of the PAIM coefficient in Figure 4.2a ($nTiO_2$) and Figure 4.2b (nZnO) showed a strong correlation of greater than 0.50 between the input variables of ZP, pH, and time to the output of HDD. Based on these results, the input of ZP, time, and pH were identified as significant input variables. On the contrary, the input variables of NOM, IS, size, and ENP concentration yielded PAIM coefficients closer to zero (< 0.20) in Figure 4.2, and Γ was less than the reference values in Table 4.2. This signified a poor relationship between these four input variables with dynamic aggregation.

Furthermore, the ZP had the highest correlation with HDD followed by pH and time in Table 4.2 (GT) and Figure 4.2 (PAIM). These findings were consistent with DLVO

theory, which explains the energy barriers, collision frequencies, and dispersions of colloidal particles in an aqueous system driven by surface charge (Hartmann et al., 2014). In addition, the interdependencies of various factors were considered to establish the possibility of synergistic and/or antagonistic effects of individual, or multiple parameters. Figure 4.2 showed only input variables of NOM, pH, and IS against ZP (as output) had a moderate correlation for both ENPs.

Table 4. 2. GT results for different input combination(s) (exclusion and inclusion shown by 0 or 1 in a mask, respectively).

Inputs	Mask	$nTiO_2$			nZnO		
		Γ	SE	V-ratio	Γ	SE	V-ratio
All	1111111	0.063	0.056	0.086	0.076	0.082	0.102
All - Conc.	0111111	0.003	0.005	0.068	0.045	0.002	0.685
All - Size	1011111	0.029	0.025	0.524	0.065	0.052	0.006
All - NOM	1101111	0.061	0.035	0.032	0.069	0.032	0.012
All - IS	1110111	0.060	0.042	0.007	0.049	0.035	0.085
All - pH	**1111011**	**0.080**	**0.120**	**0.880**	**0.103**	**0.102**	**0.156**
All -Time	**1111101**	**0.074**	**0.098**	**0.100**	**0.084**	**0.090**	**0.196**
All - ZP	**1111110**	**0.120**	**0.211**	**0.195**	**0.147**	**0.110**	**0.210**

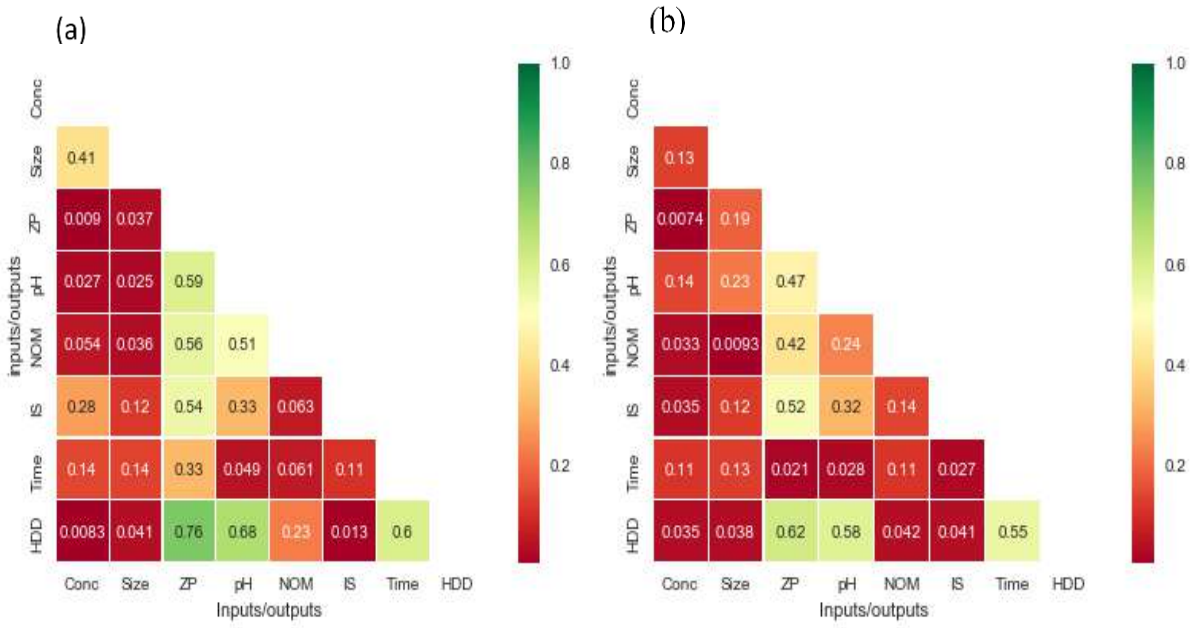


Figure 4. 2. Heat maps that depict the multidimensional interdependence among the inputs and output, based on RFPI for $nTiO_2$ (a) and nZnO (b).

4.4 Select the best combination of hyperparameters

According to Occam's Razor principle, good models can be built from the exclusion of the least significant variables (Blumer et al., 1987). To find the best optimisers for different ML, the highly ranked variables of pH, ZP, and time were assigned as the model inputs. Complete results on the training of models: ANFIS, ANN, RFR, SVR, and MLR can be found in Table A.3 ($nTiO_2$) and Table A.4 (nZnO). Taylor diagrams in Figure 4.4 ($nTiO_2$) and Figure 4.5 (nZnO) compare the performance of ML techniques on independent test datasets.

4.4.1 ANFIS

ANFIS with 3 MFs and 100 iterations showed the least MAE for both ENPs in Table 4.3. However, ANFIS models with 4 to 6 MFs had high MAE values in both training and testing data, irrespective of the number of epochs. Thus, the latter was selected as the best optimisers for developing ANFIS models. Additionally, ANFIS-based models developed using MFs of triangular (ANFIS1) and Gaussian II (ANFIS5) in Figure 4.4a ($nTiO_2$) and Figure 4.5a (nZnO), respectively, were identified as the best models for both ENPs. This was based on the highest R, lowest residual error, and the SD close to reference. The metric values for ANFIS1 and ANFIS5 were R = 0.79, RMSE = 0.10, SD = 0.10, and R = 0.70, RMSE = 0.20, SD = 0.19, respectively. ANFIS models based on Gaussian I MF (ANFIS4) had the lowest performance for both datasets. The results herein contrast previous observations that demonstrated the Gaussian function had higher performance and ability to smoothen non-linear data relative to other MFs (Tan et al., 2017). Therefore, the choice of MF-type is problem and data-dependent rather than intuitive.

Table 4. 3. Optimization process of ANFIS using MF and Epochs

	MAE(nZnO)		**MAE($nTiO_2$)**	
Model	**Train**	**Test**	**Train**	**Test**
ANFIS 3 MF 50 Epochs	0.17	0.18	0.19	0.21
ANFIS 3 MF 100 Epochs	0.12	0.11	0.16	0.19
ANFIS 3 MF 600 Epochs	0.12	0.16	0.17	0.20
ANFIS 4 MF 50 Epochs	0.14	0.15	0.16	0.22
ANFIS 4 MF 100 Epochs	0.15	0.16	0.17	0.21

ANFIS 4 MF 600 Epochs	0.12	0.18	0.17	0.23
ANFIS 5 MF 50 Epochs	0.18	0.21	0.20	0.23
ANFIS 5 MF 100 Epochs	0.17	0.21	0.21	0.23
ANFIS 5 MF 600 Epochs	0.13	0.19	0.21	0.22

4.4.2 ANN

In Figure 4.3a for $nTiO_2$ (MAE_{train} = 0.03 and MAE_{test} = 0.05), and Figure 4.3b for nZnO (MAE_{test} = 0.06 and MAE_{train} = 0.08) the ANN with 10 neurons had the least MAE. This was the optimal network to build ANN models. Furthermore, the combination of adaptive momentum (Adam) and hyperbolic tangent (tanh) functions (ANN1) in Figure 4.4b ($nTiO_2$) showed high performance with metric values of R = 0.83, RMSE = 0.09, and SD = 0.11. In Figure 4.5b (nZnO), the models with the activation functions of the hyperbolic tangent (ANN1) and the sigmoid (ANN2) had an equal correlation with a value of R = 0.90. However, ANN1 showed low RMSE and SD close to the “reference SD” (RMSE = 0.11, SD = 0.19) compared to ANN2 (RMSE = 0.12, SD = 0.18), therefore, was selected as the best combination.

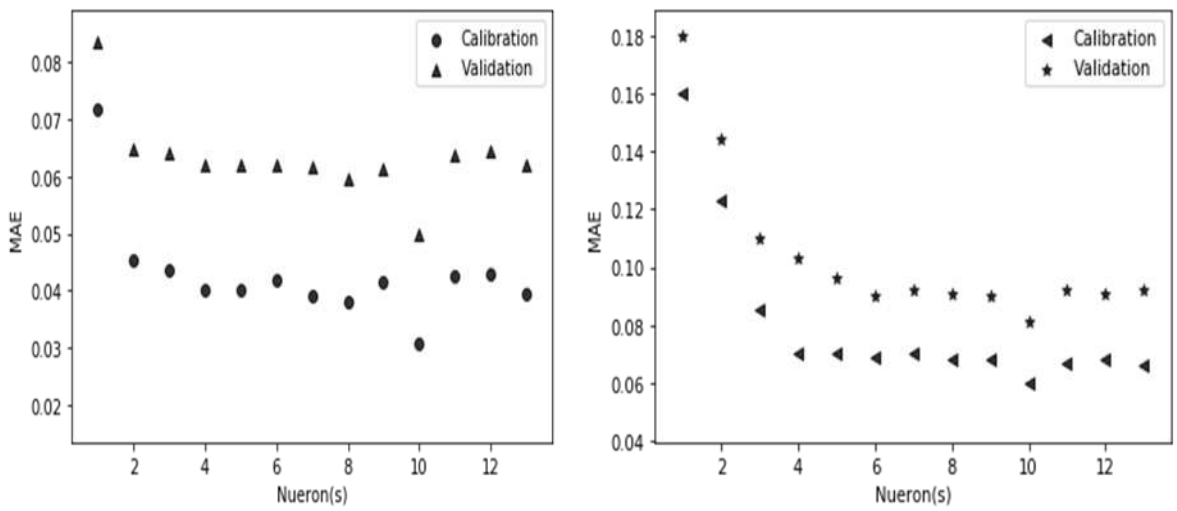


Figure 4. 3. Model performance based on the number of neurons for ANN for (a) $nTiO_2$ data and (b) nZnO

Overall, the ANN1-3 models based on the Adam showed higher performance compared to ANN3-6 models that used the stochastic gradient descent (SGD) optimisers. The high performance of the former was due to its adaptive high learning rate, and ability to deal with noise, unlike the gradient descent which has stagnation

at the local minima of the curve (Zarra et al., 2019). Moreover, ANN models using the activation function of rectified linear units (ReLU) have been defined as computationally efficient due to an expanded range of [0,10] (Rynkiewicz, 2019). However, the results herein demonstrated high predictive accuracy for the models based on the tanh activation function. Therefore, the choice of a transfer function is problem-dependent.

4.4.3 RFR

In Figure 4.4c ($nTiO_2$) and Figure 4.5c (nZnO), the RFR models generated high prediction accuracy with R > 0.90 for both ENPs. The combination of 20 trees and 42 randomised states (RFR1) in Figure 4.4c ($nTiO_2$), was identified as the best model (R = 0.93, RMSE = 0.07, SD = 0.12). Increasing the decision trees from 20 to 100 decreased the accuracy of the models. This was in contrast to previously reported results where an increase in decision trees corresponded to a reduction in overfitting, and an increase in accuracy (Hou et al., 2020). The observed contradiction was plausibly due to the small database used in this study. To account for these results, previous studies (Oshiro et al., 2012; Probst and Boulesteix, 2017) indicate that the error rate at times can be a non-monotonous function of the number of trees. Thus, fewer trees can produce better performance compared to large trees for small numerical datasets. Moreover, in Figure 4.5c (nZnO) the three RFR models had equal performance, i.e.: (R = 0.91, RMSE = 0.11, and SD = 0.19); therefore, for convenience, RFR1 was chosen.

4.4.4 SVR

Results in Figure 4.4d ($nTiO_2$) and Figure 4.5d (nZnO) showed the best model was SVR3 (C (penalty coefficient) = 1, ε (epsilon) = 0.1, and γ (gamma) = 10) for both ENPs. An increase of ε from 0.1 to 0.3 produced a reduction in R; while an increase in γ values from 0.1 to 10 resulted in improved model performance. SVR3 model had a low residual error with metric values of (R = 0.85, RMSE = 0.10, SD = 0.11) and (R = 0.89, RMSE = 0.12, SD = 0.18) in Figure 4.4d ($nTiO_2$) and Figure 4.5d (nZnO), respectively. SVR4-6 models that used a kernel function of a polynomial (poly) had the least performance outcomes compared to the SVR1-3 models based on the radial basis function (RBF).

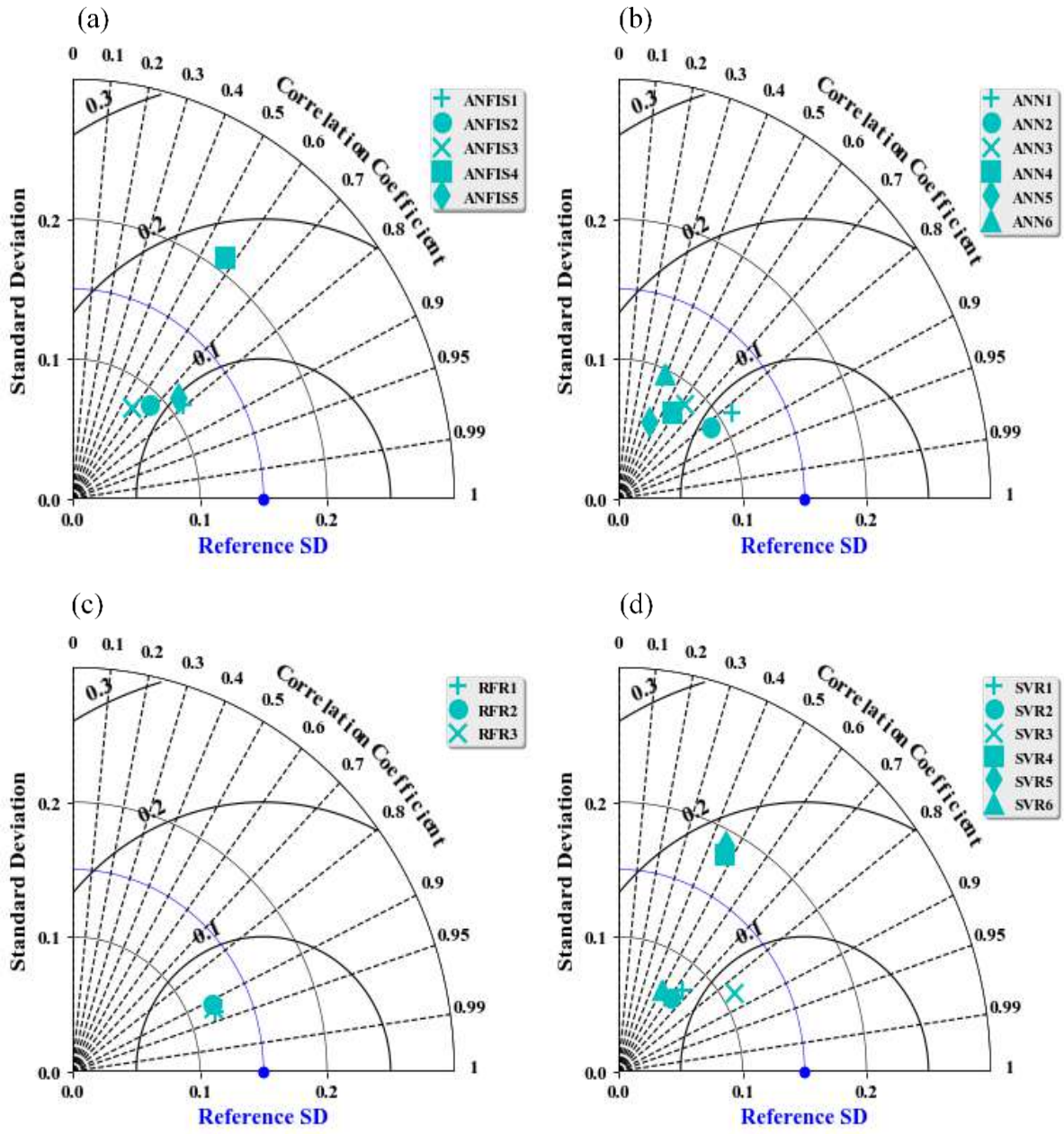


Figure 4. 4. Taylor diagram comparing the performance of ML models (a) ANFIS, (b) ANN, (c) RFR, and (d) SVR developed using the $nTiO_2$ dataset.

In Figure 4.4, the blue contour at a radial distance of 0.1, depicted as “Reference SD” is the standard deviation for testing datasets. The dark black contour and dashed lines represent the centered RMSE regimes, and the R values, respectively. ANFIS models (ANFIS1 – triangular MF, ANFIS2 – trapezoidal MF, ANFIS3 – generalized bell MF, ANFIS4 – Gaussian-I MF, ANFIS5 – Gaussian-II MF), ANN based on Adam optimizer (ANN1 – tanh, ANN2 – sigmoid, ANN3 – rectified linear unit), ANN based on gradient descent optimizer (ANN4-tanh, ANN5 – sigmoid, ANN6 – rectified linear unit), RFR based on 42 randomised states (RFR1 – 20 decision trees, RFR2 – 60 decision trees,

RFR3 – 100 decision trees), and SVR based on radial basis function (SVR1 – (C = 1, ε = 0.1, and γ = 1), SVR2 – (C = 1, ε = 0.3, and γ = 1), SVR3 – (C = 1, ε = 0.1, and γ = 10)), SVR based on polynomial function (SVR4 – (C = 1, ε = 0.1, and γ = 1), SVR5 – (C = 1, ε = 0.3, and γ = 1), SVR6 – (C = 1, ε = 0.1, and γ = 10)).

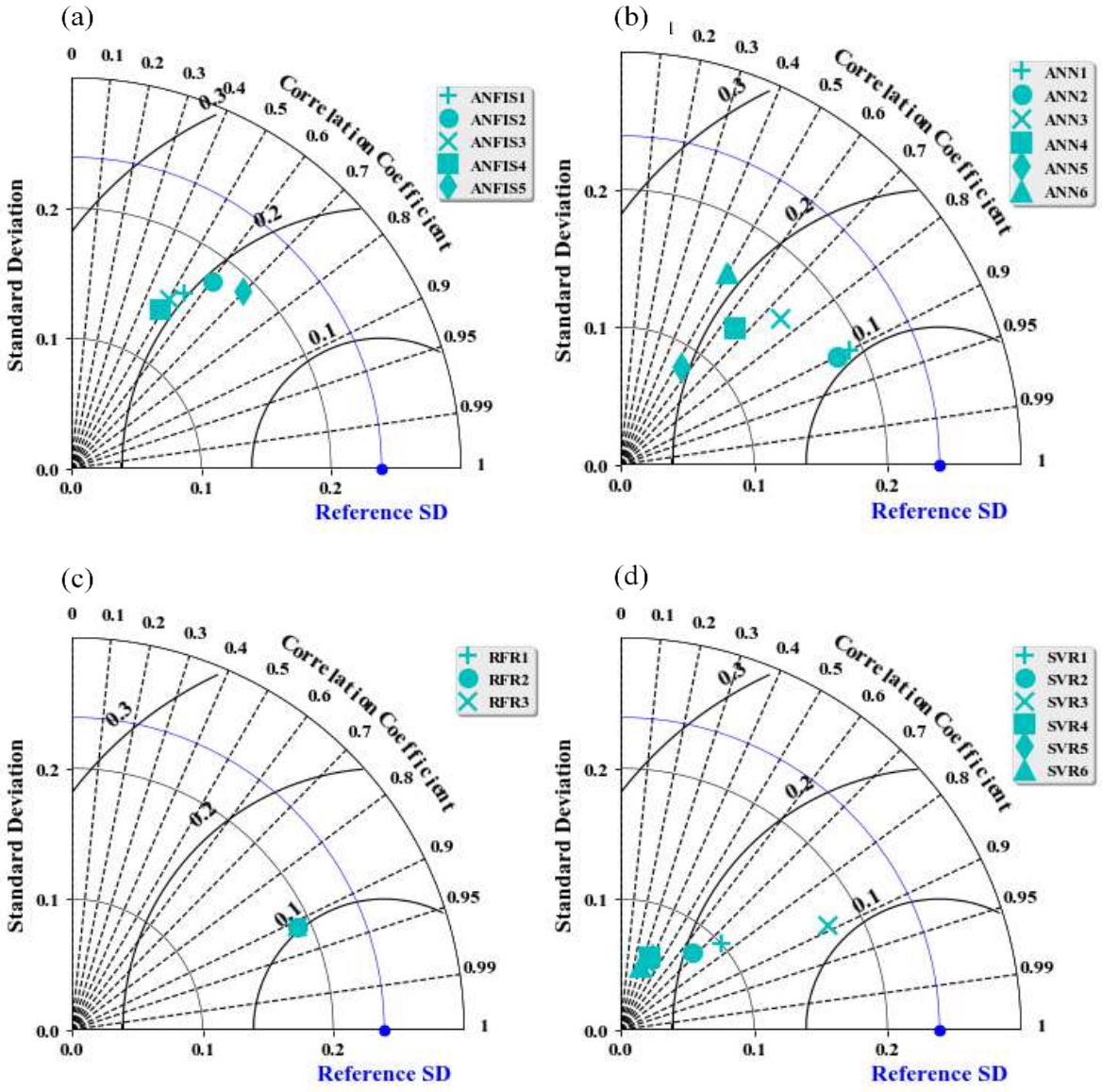


Figure 4. 5. Taylor diagram comparing the performance of ML models (a) ANFIS, (b) ANN (c) RFR, and (d) SVR developed using the nZnO dataset.

In Figure 4.5 the blue contour at a radial distance of 0.24 depicted as “Reference SD” is the standard deviation for testing datasets. The dark black contour and dashed lines represent the centered RMSE regimes and the R values, respectively. The description of parameters for each ML technique is the same as in Figure 4.4. Results show consistency with previous literature studies, as the RBF kernel function generally

shows better performance than other kernel functions. For example, the RBF-SVM and polynomial-SVM models, showed an NSE of 0.68 and 0.65, respectively for estimating river-suspended sediments (Choubin et al., 2018). For estimating the total organic carbon content the Gaussian, linear, and polynomial functions showed correlation coefficients (R) of 0.7792, 0.6958, and 0.7129, respectively (Rui et al., 2019). Additionally, the RBF function displayed an R^2 value of 0.9902 and the linear function an R^2 value of 0.9702 for investigating the parameters effective on the performance of a humidification-dehumidification seawater greenhouse (Zarei et al., 2018). The high performance of models using the RBF kernel function was attributed to the approach's suitability for the linear inseparable problem(s); whereas the polynomial kernel performs better for orthogonal normalised data (Zarei et al., 2018).

4.5 Comparing the performance of ML models

The best models for each ML technique using Taylor diagrams were used to compare the performance of developed models. These were ANFIS1, ANN1, RFR1, and SVR3 for the $nTiO_2$ dataset, and ANFIS5, ANN1, RFR1, and SVR3 for nZnO. The performance of these ML techniques was compared based on their ability to accurately predict the distribution or characteristics of independent data not used during training. Figure 4.6 and Figure 4.7 show that RFR1 had the highest predictive performance and MLR the least. For example, in Figure 4.6c for the $nTiO_2$ dataset, the RFR1 shows high R^2 and low RMSE of 0.81 and 0.08, respectively.

Conversely, the MLR yielded the least performance with a low R^2 of 0.02 and a high RMSE of 0.15 in Figure 4.6e. Similarly, for the nZnO dataset in Figure 4.7c, the RFR1 had high R^2 and NSE values of 0.83 and 0.80, respectively. The ML models of RFR1, SVR3, and ANN1 demonstrated satisfactory performance using the rating systems of R2 and NSE for both ENPs. However, ANFIS and MLR models showed relatively large RMSE and NSE of less than 0.40 rated as unsatisfactory for both ENPs. This indicated their unsuitability for application in developing a decision-supporting tool to predict the dynamic aggregation of ENPs based on data sets used in this investigation.

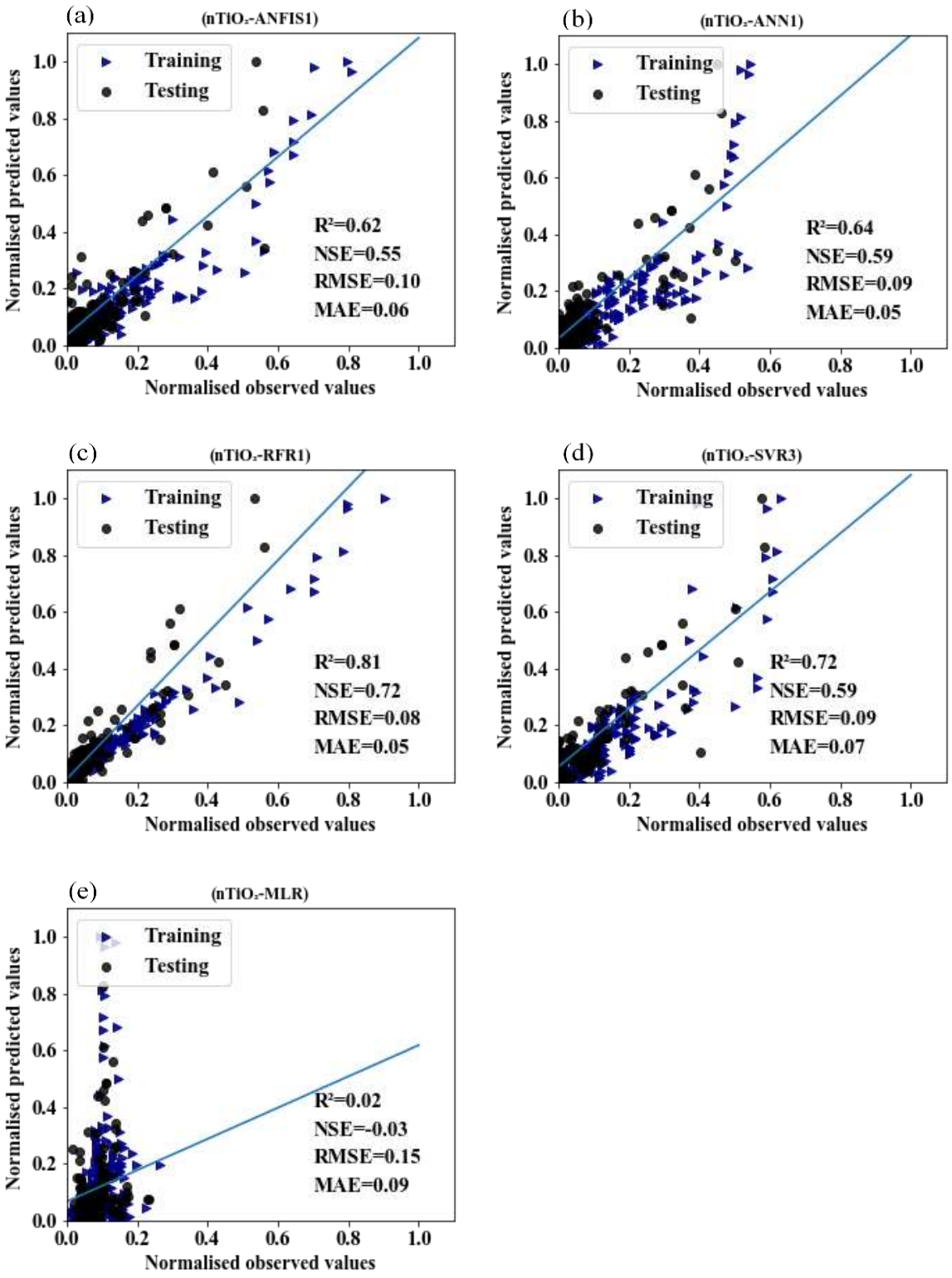


Figure 4. 6. Scatter plots for the predicted models derived from $nTiO_2$ data using high-ranked input variables of (pH, ZP, and time): (a) ANFIS1, (b) ANN1, (c) RFR1, (d) SVR3, and (e) MLR.

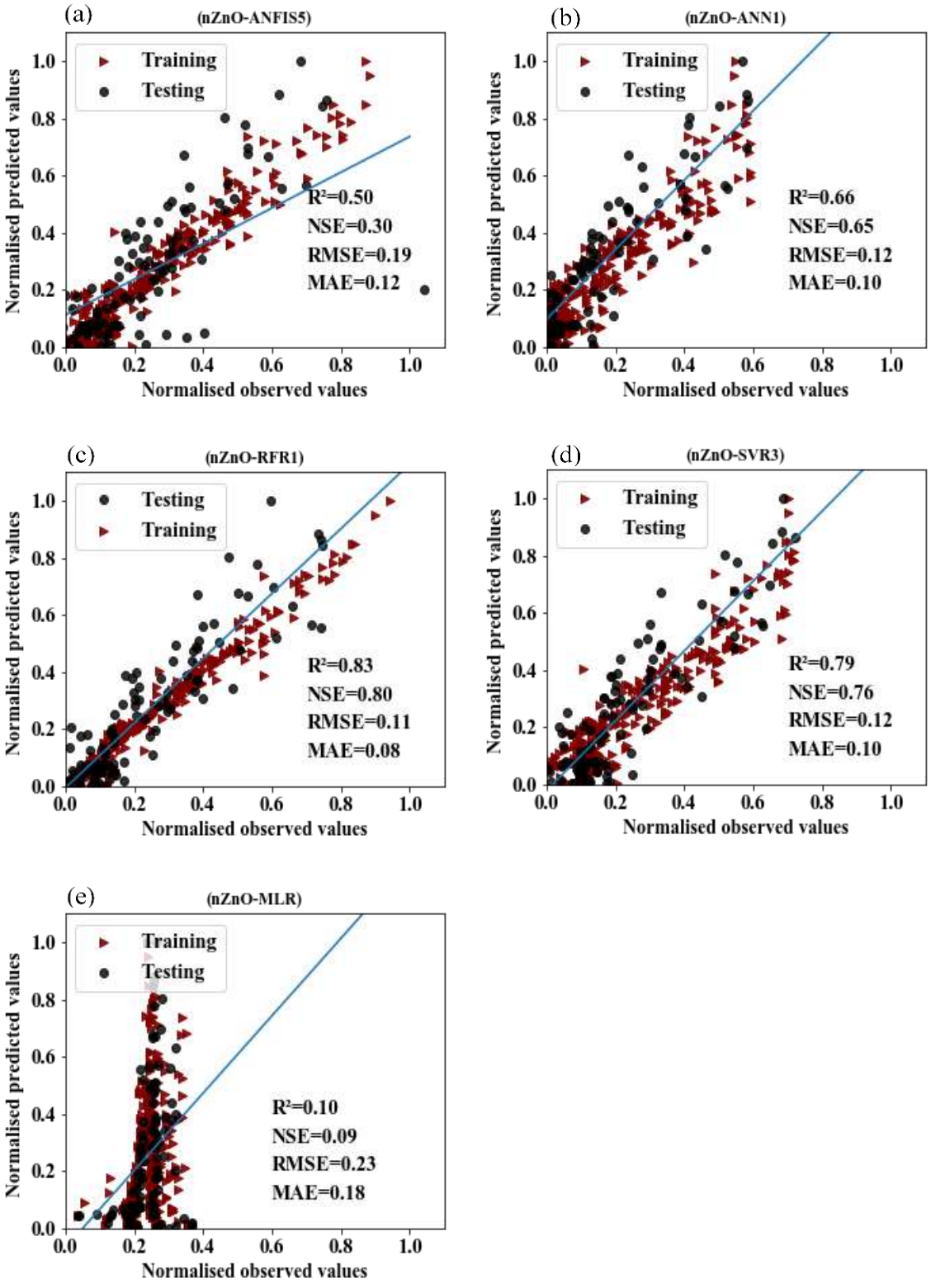


Figure 4. 7. Scatter plots for the predicted models derived from nZnO data using high-ranked input variables of (pH, ZP, and time): (a) ANFIS5, (b) ANN1, (c) RFR1, (d) SVR3, and (e) MLR.

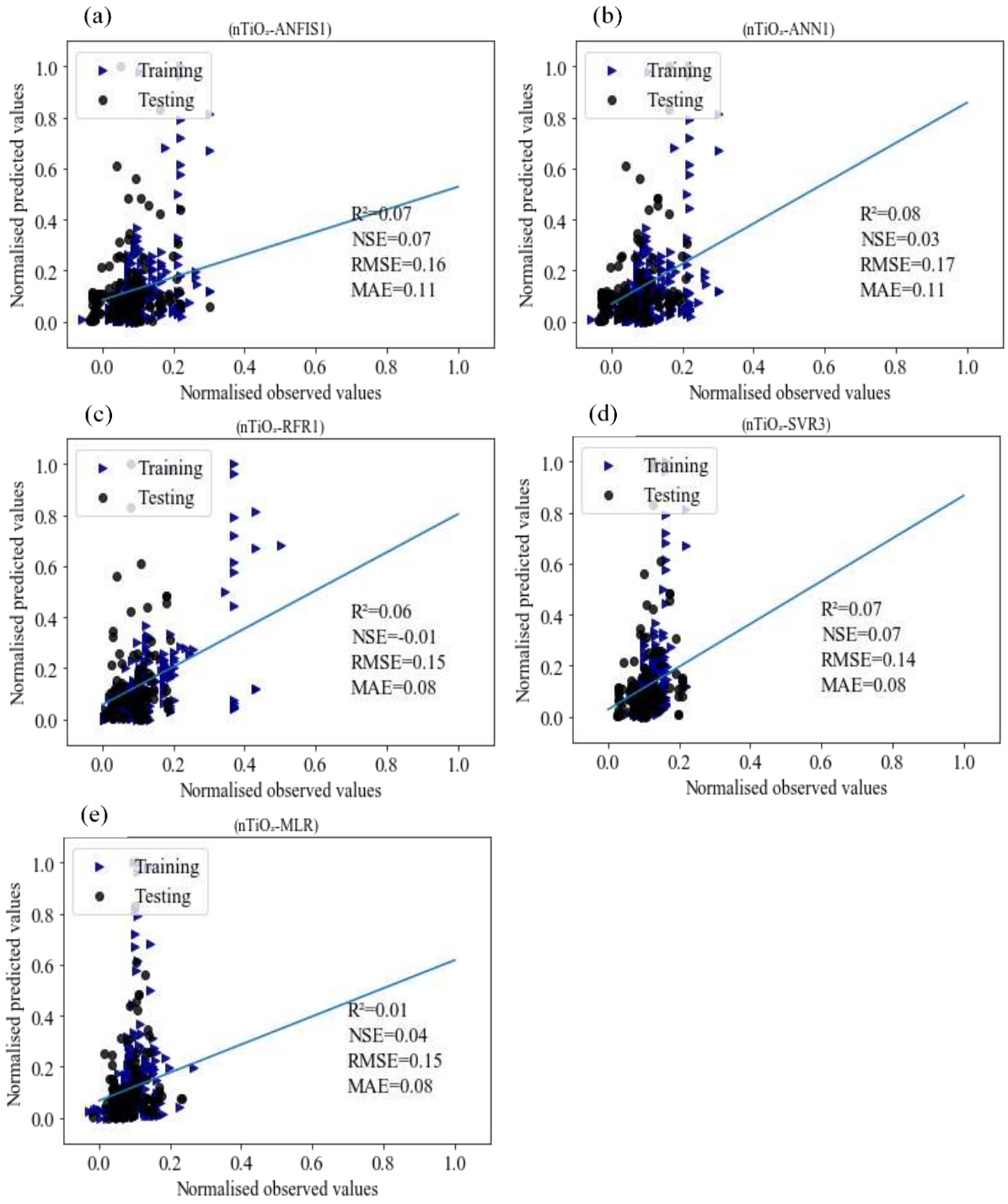


Figure 4. 8. Scatter plots depicting predicted models derived from the $nTiO_2$ data set using low-ranked input variables of NOM, IS, ENP concentration and size: (a) ANFIS1, (b) ANN1, (c) RFR1, (d) SVR3, (e) MLR for $nTiO_2$.

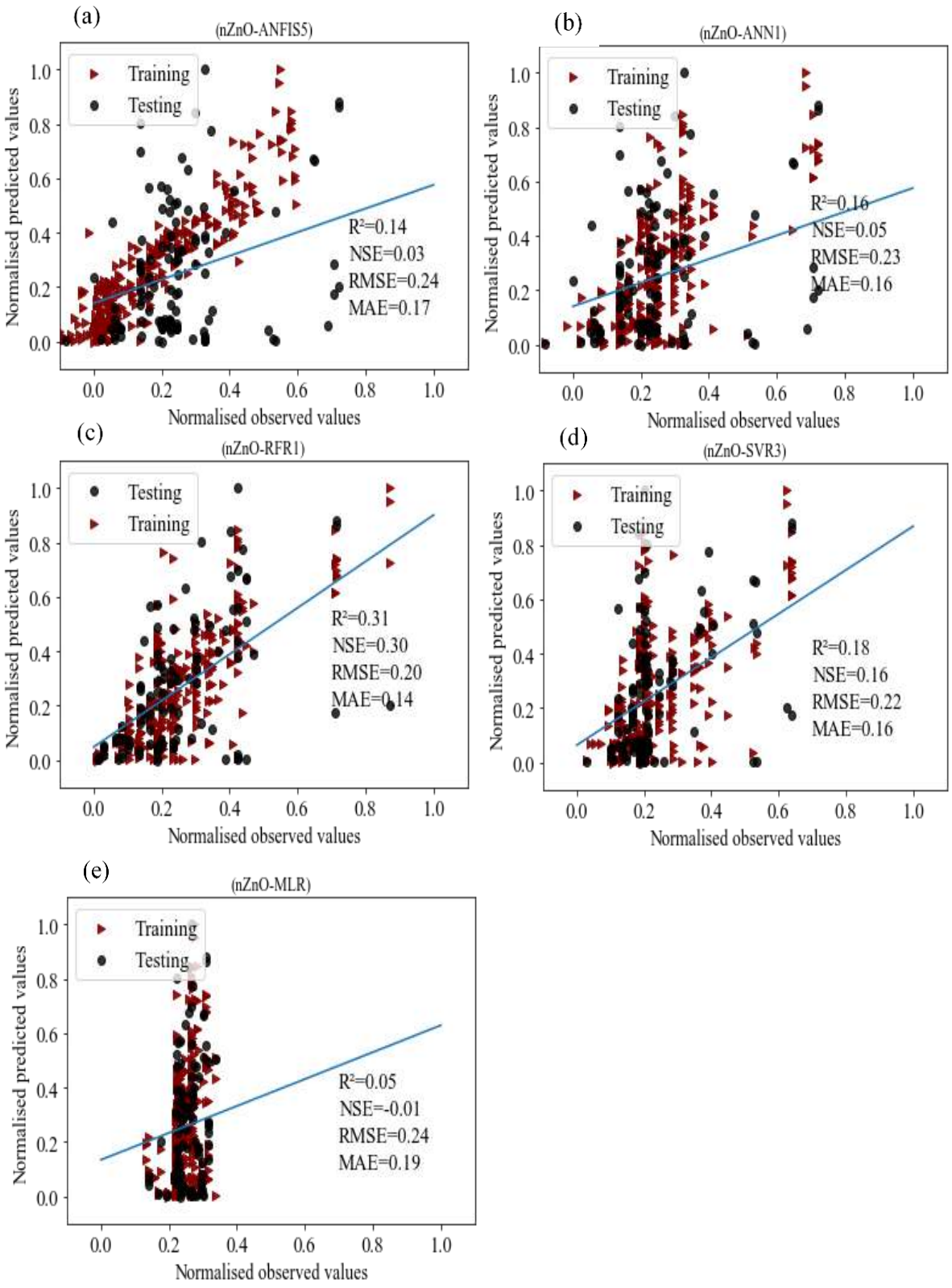


Figure 4. 9. Scatter plots depicting predicted models derived from the nZnO data set using low-ranked input variables of NOM, IS, ENP concentration and size: (a) ANFIS5, (b) ANN1, (c) RFR1, (d) SVR3 and (e) MLR.

In addition, the violin plots in Figure 4.10a ($nTiO_2$) and Figure 4.10b (nZnO) were used to visualise the distribution of predicted values against the independent data sets. In Figure 4.10, the white dots on violin plots depict the mean of each dataset with a wide regime signifying high probability distribution, whereas in the skinner regime, the converse holds. The boxes bound IQR (25^{th}, 50^{th}, and 75^{th} quartiles). The ends of solid black points depict the highest (95^{th}) and lowest values (5^{th}). The ML models of RFR1, SVR3, and ANN1 showed good fit at the interquartile range (IQR) range and poor fit at the 95th percentile plausibly, due to overfitting during learning, or the existence of outliers in the latter (Wang et al., 2019). The high performance of RFR was attributed to its non-parametric nature, robustness against overfitting, and high error tolerance (Wang et al., 2019). On the contrary, the poor prediction using the MLR in both datasets was related to the technique's inability to handle data that have no predefined linear correlations, between input and output variables (Aquilina et al., 2018).

The results of ML models derived using the input valuables, viz. NOM, size, IS, and ENP concentration shown in Figure 4.8 ($nTiO_2$) and Figure 4.9 (nZnO) demonstrated poor and unsatisfactory prediction performance with statistical metrics of R^2 and NSE < 0.4 (below acceptable threshold). In addition, the violin plots in Figure 4.10c ($nTiO_2$) and Figure 4.10d (nZnO) showed poor distribution for the predicted values compared to the test data for all developed ML models. In this study, the ZP, the pH, and time were found to be highly suitable parameters for developing a parsimonious model to predict the levels of aggregation of ENPs in freshwater-like exposure media. Generally in the literature, a ZP of greater than or equal to ± 30 mV indicates stable ENP with high electrophoretic mobility and, in turn, exhibits high colloidal dispersion in aqueous media (Hartmann et al., 2014; Lowry et al., 2016). Additionally, as the pH of the media shifts towards the point of zero charge (PZC) or isoelectric point (IEP), the surface charge of the ENPs decreases correspondingly, resulting in the highest aggregation (Amde et al., 2017; Loosli et al., 2014; Peng et al., 2015).

However, these findings appeared to contradict the previously reported results as NOM and IS, etc., were generally considered to play a significant role in the transformation processes of ENPs in aquatic environments (Abbas et al., 2020; Amde et al., 2017; Philippe and Schaumann, 2014). For example, IS has been shown to control the aggregation of ENPs by compressing the electric double layer (EDL) and, in turn, reducing the double length, thus increasing the aggregation (Amde et al., 2017; Peng et al., 2015). NOM has been demonstrated to play a significant role in inhibiting

(by steric stabilisation at low IS) or promoting (by a chelating effect or bridging at high IS) the aggregation of ENPs (Abbas et al., 2020; Leareng et al., 2020).

The observed difference(s) in predominant influencing factors can be accounted for two-fold; first, by using the prediction *versus* significance concept (Lo et al., 2015). The concept indicates that momentous input variables in a given domain do not always imply stronger prediction variables, especially for non-linear complex problems (Van't Veer et al., 2002; Welch and Goyal, 2008). For example, Van't Veer et al. (2002) demonstrated that the highly regarded influential variables of *lymph node status* and *histological grae* were poor model inputs for the classification of breast cancer. Similarly, variables found to be significant for fluctuations in the stock market index were shown to have no predictive power (Welch and Goyal, 2008). Lo and colleagues (Lo et al., 2015) also pointed out that key factors for good prediction *versus* significant variables are not easily discernible, but rather are dependent on different properties of the underlying distribution for each parameter in question.

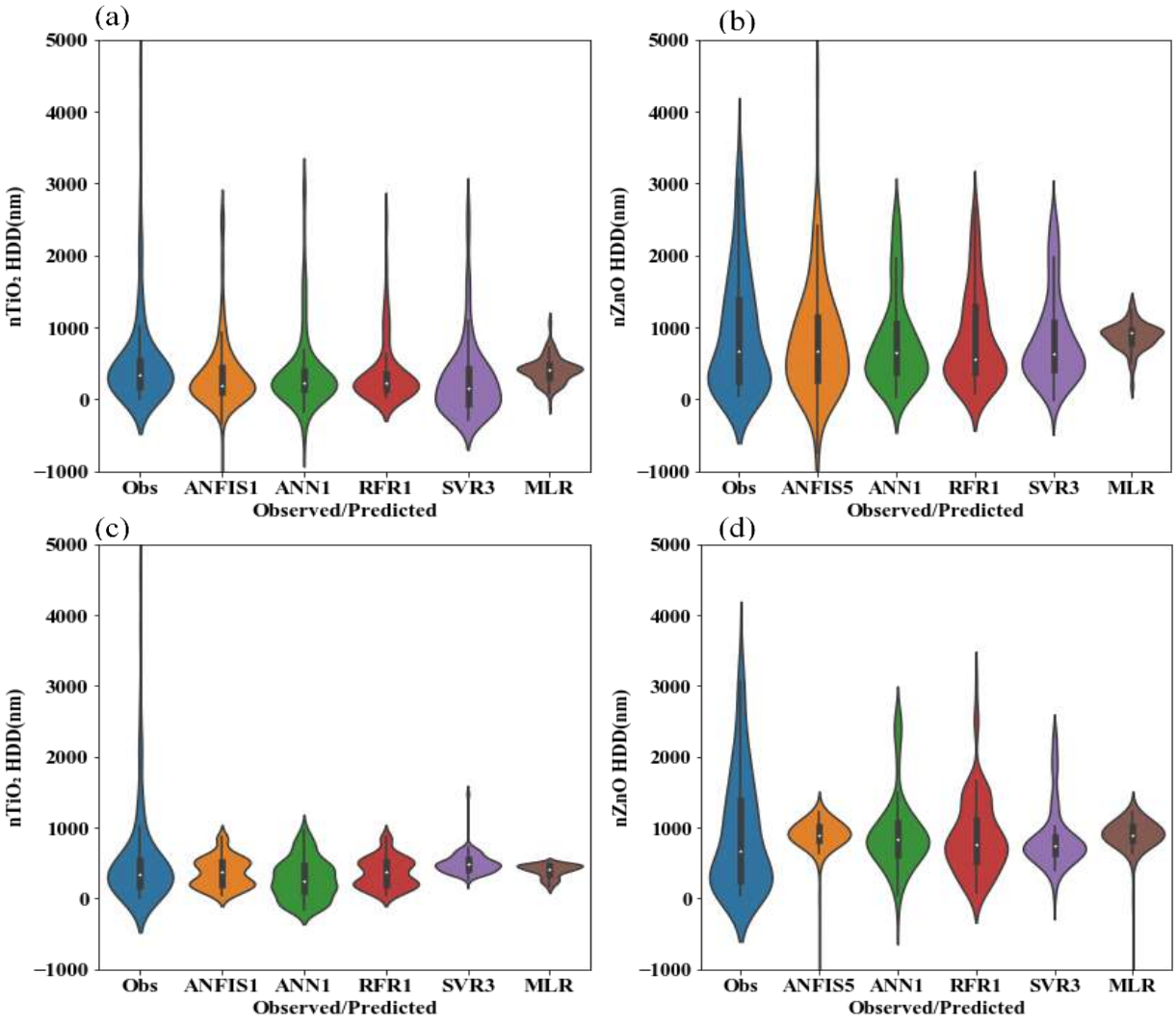


Figure 4. 10. Density mass distribution of predicted values against the observed (Obs) using violin plots to compare the model performance. Models (a) and (b) are based on high-ranked, (c) and (d) low-ranked variables on the aggregation of $nTiO_2$ and nZnO, correspondingly.

Second, the poor prediction results based on NOM and IS as model inputs can be attributed to their high dependence on other underlying controlling factors and distribution. For example, in many studies, NOM is generally treated as a single group without general consideration of different heterogeneity (Louie et al., 2013). However, various types of NOM found in natural systems that include high molecular weight (MW) (humic acid, fulvic acid, etc.), and low MW (carboxylic, amines, etc.) organic materials have different MW distribution, polydispersity, and chemical properties which may influence the aggregation size and fractal dimension of ENPs differently (Abbas et al., 2020; Louie et al., 2016, 2013). In addition, the IS depends on numerous aspects such as solution pH, electrolyte type, concentration, etc. (Abbas et al., 2020; Chowdhury et al., 2012a; Peng et al., 2017a). For example, the effect of divalent electrolytes such as calcium chloride ($CaCl_2$) has been demonstrated to be three-fold higher compared to monovalent electrolytes, such as sodium chloride ($CaCl_2$) at the same molar concentration (Chowdhury et al., 2012a).

Additionally, based on the applied evidence-based approach (Tolaymat et al., 2017), the available data considered all $nTiO_2$ as a single group of ENPs; yet $nTiO_2$-rutile and $nTiO_2$-anatase are known to exhibit different aggregation profiles (Liu et al., 2011). Also, ENPs investigated in many works were uncoated; however, different types of coating impact the aggregation differently (Ellis et al., 2016). Therefore, despite the ML in this study, demonstrating high prediction accuracy, the limitations can include data outside the trained and tested herein (Kovalishyn et al., 2018). As the data become more accessible this raises the need for additional inherent ENP features and consideration of other types or classes not identified in this work. This is because refined model resolution can improve the applicability and robustness of the developed models.

4.6 Chapter summary

In this chapter, supervised ML algorithms of RFR, SVR, and ANN were successfully applied to deduce the underlying patterns and predict the dynamic aggregation of

ENPs in aqueous systems using data solicited from diverse studies documented in the literature. The input variables namely; the ZP, pH and time showed high predictive strength, and therefore, should be prioritized in future experimental investigations as indicative precursor conditions for rapid initial screening and the development of robust nano-safety frameworks to optimise societal benefits and for long-term proactive ecological protection. ML algorithms developed herein were able to handle heterogeneous data and are expected to inspire further systematic interrogation of key controlling parameters for other ENP transformation processes (e.g., dissolution, adsorption, deposition, etc.). This can aid in supporting effective policy formulation and reduce the cost and laboratory work associated with experimental testing that cannot be easily manageable.

Chapter 5. Developing machine learning for predicting the dissolution of zinc oxide nanoparticles in aqueous systems

This chapter describes the results ofa range of ML algorithms (SVR, ANN, MLR, RFR, and XGB) to elucidate the underlying PC and WC properties and develop predictive algorithms for dissolution of nZnO in aquatic environments.

5.1 Introduction

The dissolution of nZnO is considered one of the important chemical processes that impact biodurability, bioavailability, bioaccumulation, and possible toxicological effects (Hou et al., 2018; Mahaye et al., 2017; Musee et al., 2014). Generally, 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 reflected; despite these interactions have a significant impact on toxicity.

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 chapter describes the results of the use of data-driven methods such as ML to provide a clear understanding of the interactions and mechanisms underlying the dissolution of ENPs as the volume of experimental data increases.

5.2 Analysis of heterogeneous data on nZnO dissolution

Data extraction of nZnO dissolution in aquatic systems was performed using the process described in Sections 3.1 - 3.3. A total of 791 data points were extracted on nZnO dissolution from publications that are summarised in Table B.1. The dataset initially had continuous (*n=7*) and categorical (*n=5*) features namely; ENP 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 ZP (X_{12}, mv). The concentration of Zn^{2+} (X_{13}, mg/ℓ) was used as a model output. The input variables of IS, shape, ZP, NOM, nZnO concentration, and time had missing values as indicated in Table 5.1. The KNN imputation method with a value of k = 9 was used to compute the missing values. Furthermore, shape and ZP had insufficient data and these inputs were removed, as they were likely to result in data snooping. In addition, the categorical variables were converted to 1 and 0 using **Algorithm 1.**

Table 5. 1. Type of data with missing percentages, and descriptive statistics for input variables

Variables	Units	Type of inputs	Missing data (%)	Mean	SD
NOM	$\mathrm{mg}\cdot\ell^{-1}$	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	$\mathrm{mg}\cdot\ell^{-1}$	Continuous	1.07	0.19	0.26
Time	Hour	Continuous	2.03	0.11	0.24
Shape	–	Categorical	50.3	–	–
ZP	mV	Continuous	62.4	0.12	0.17
NOM type	–	Categorical	–	–	–
Coating	–	Categorical	–	–	–
Coating type	–	Categorical	–	–	–
Salt type		Categorical	–	–	–
Zn^{2+} concentration	$\mathrm{mg}\cdot\ell^{-1}$	Continuous	–	0.24	0.21

5.3 Feature correlation analysis

Variable selection helps avoid the dimensionality curse, reduces bias or noise, and improves model generalisation. The results described in Figure 5.1 did not show existing multi-collinearity as the vif scores were less than 5. In addition, PAIM and XGBoostFI results in Figure 5.2 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.

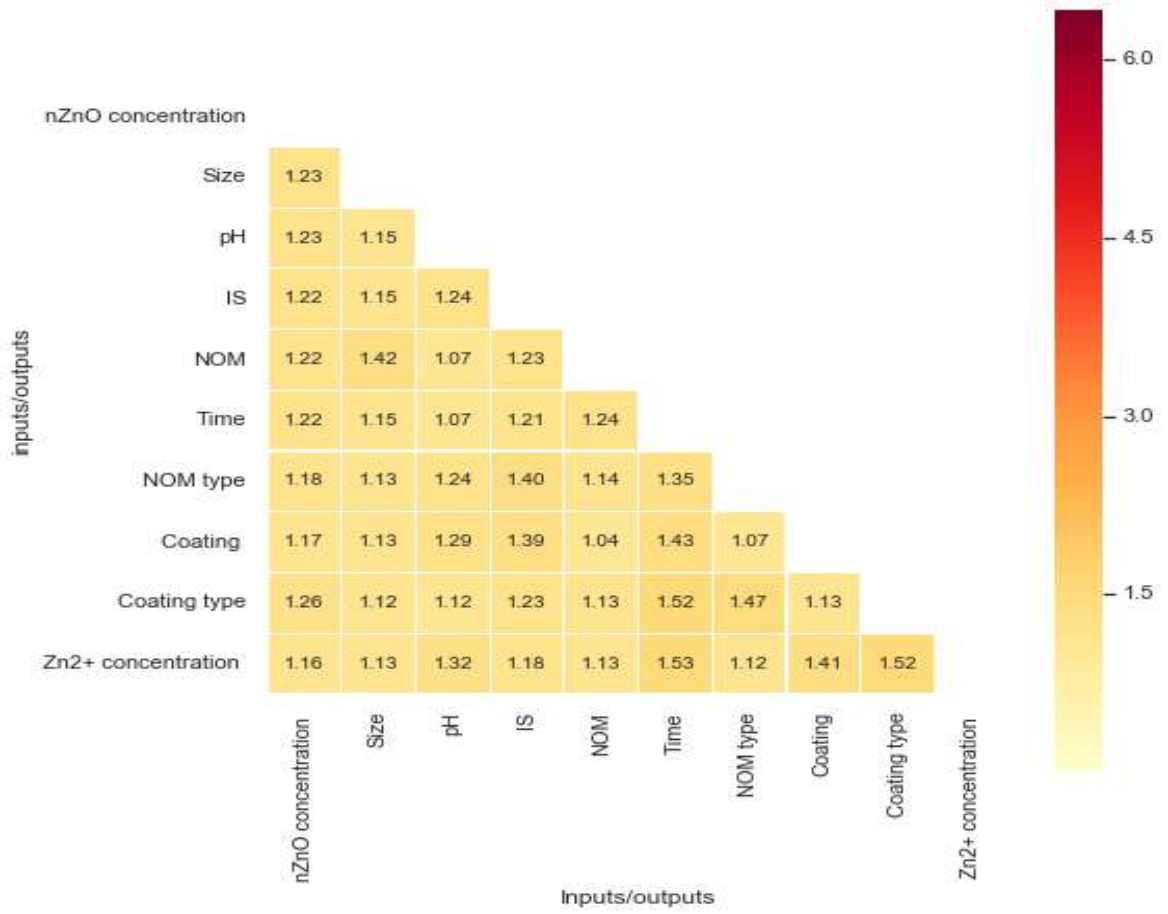


Figure 5. 1. Vif values to estimate the multi-collinearity

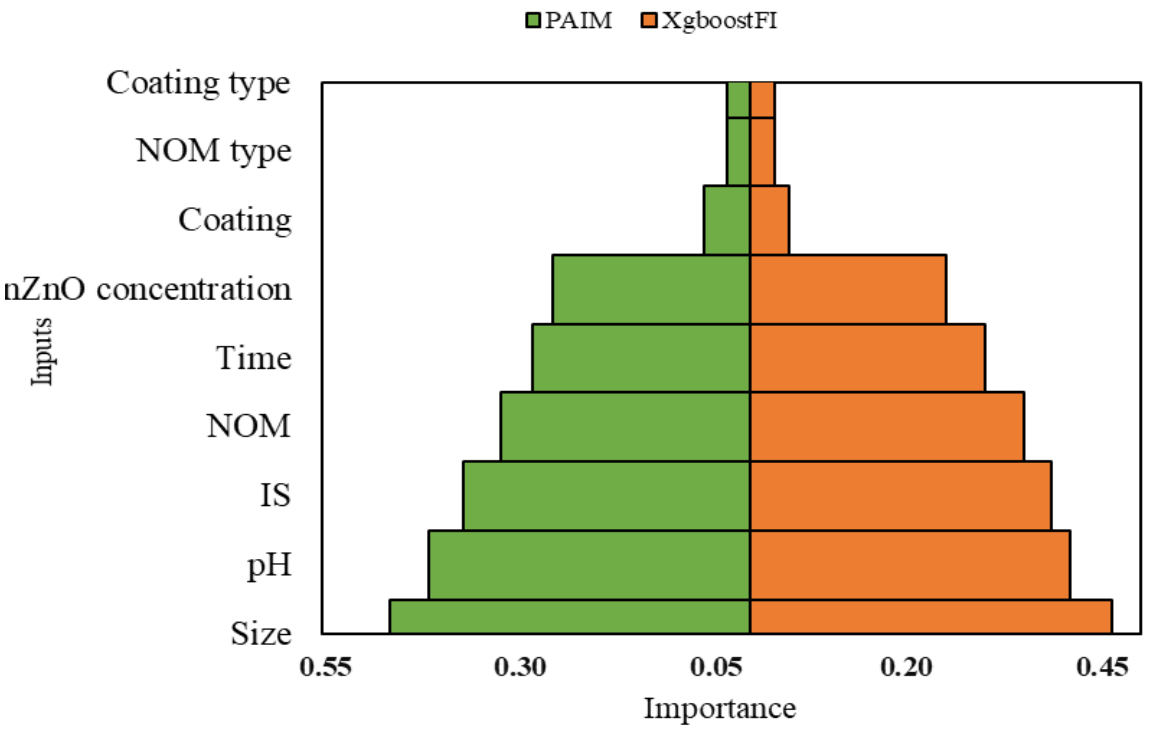


Figure 5. 2. Bipyramid depicting both PAIM and XGBoostFI results

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., 2012a). 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 Figure 5.2 of categorical variables the concept of causal vs 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). Therefore, as covered in our previous study (Yalezo and Musee, 2023), significant variables are not always good predictor variables. Prediction is influenced by correlation as opposed to causal effect; hence, at times experimental reported significant variables do not automatically possess high predictive power.

5.4 Selecting optimisers for different ML algorithms

5.4.1 ANN

Results in Table 5.2 were developed based on the ANN architecture with 11 neurons, which had the least MAE in Figure 5.3a. The best model was the ANN3 model 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 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 (Rahman et al., 2015).

Sigmoid functions often have a range less than an absolute unit and a comparable efficiency. The sigmoid-based function exhibits constant, linear, and curved behaviors, and it is also smoothly differentiable (Chen et al., 2015; Rojas, 1996). Nevertheless, their narrow range negates gradients when they are saturated, and their outputs are not zero-centered (Pushpa and Manimala, 2014). On the other hand, non-sigmoid functions, including ReLU and its derivatives, Leaky ReLU, and ELU, have non-saturation in the positive regime, are highly computationally efficient, and in reality converge significantly quicker than sigmoid/tanh (Clevert et al., 2015; Goodfellow et al., 2013; Pham et al., 2019).

Further, in practice, GD based approach shows a good fit for ANN with a large number of parameters. However, sluggish convergence and stagnation at the local curve minima are some of the disadvantages (Rahman et al., 2015). These features are linked to the position of the gradient vector or learning process that is static for all weights (Burney et al., 2007; Rehman and Nawi, 2011). As a result, the accuracy and convergence rate of traditional-based GD have been enhanced by the introduction of gradient descent-based optimizations such as the adaptive momentum (Adam) (Kingma and Ba, 2014), AdaGrad (Duchi et al., 2011), and RMSProp (Tieleman and Hinton, 2012). Attributes of Adam are computational efficiency, reduced memory usage, ability to handle complex problems with noisy and sparse slopes, intuitive

interpretation of hyper-parameters, minimal optimization required, and ability to update the learning rate as the learning process progresses (Duchi et al., 2011). Compared to other conventional GD approaches, this method is receiving significant consideration for deep learning applications, especially for image recognition (Hameed et al., 2016; Rehman and Nawi, 2011; Shao and Zheng, 2011).

5.4.2 RFR

To improve the model accuracy, it is common to build RF models using a large number of trees (typically between 100-1000 or higher) (Hou et al., 2020; Liaw and Wiener, 2002). However, in Table 5.2 the RFR with 100 trees had the highest performance with R = 0.92, and RMSE = 0.06. The model performed admirably when the number of trees increased from 20 to 100 and further increase from 200 to 500 trees resulted in poor performance. Increasing trees beyond the threshold of saturation does not inherently enhance accuracy, especially for small numerical data sets (Oshiro et al., 2012)

5.4.3 SVR

In Table 5.2 the SVR2 model 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 radial basis function (RBF) models performed better than polynomial (poly). This was in agreement with other previously published results (Papa et al., 2015; Subramanian and Natarajan, 2021; Yalezo and Musee, 2023).

5.4.4 XGB

The accuracy of XGBoost models is rooted in a large variety of hyper-parameters such as learning rate, maximum depth, regularization, or penalty term on weights among others. The learning rate is used to prevent overfitting by making the boosting process more conservative. Doing so reduces the influence of each tree and leaves space for future trees to improve the model. A low value means that the model is more robust to overfitting. The intuition behind this technique is that it is better to improve a model by taking many small steps than a few large steps (Natekin & Knoll, 2013). Moreover, the

maximum depth or size of a tree is the number of splits in each tree. Maximum depth controls the complexity of the boosted structure.

The XGBoost model is based on the concept of ensemble learning similar to that of RF. Results in Figure 5.3b and Figure 5.3c showed the XGBoost models reached the highest performance with a learning rate of 0.10 and max 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 5.2 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.

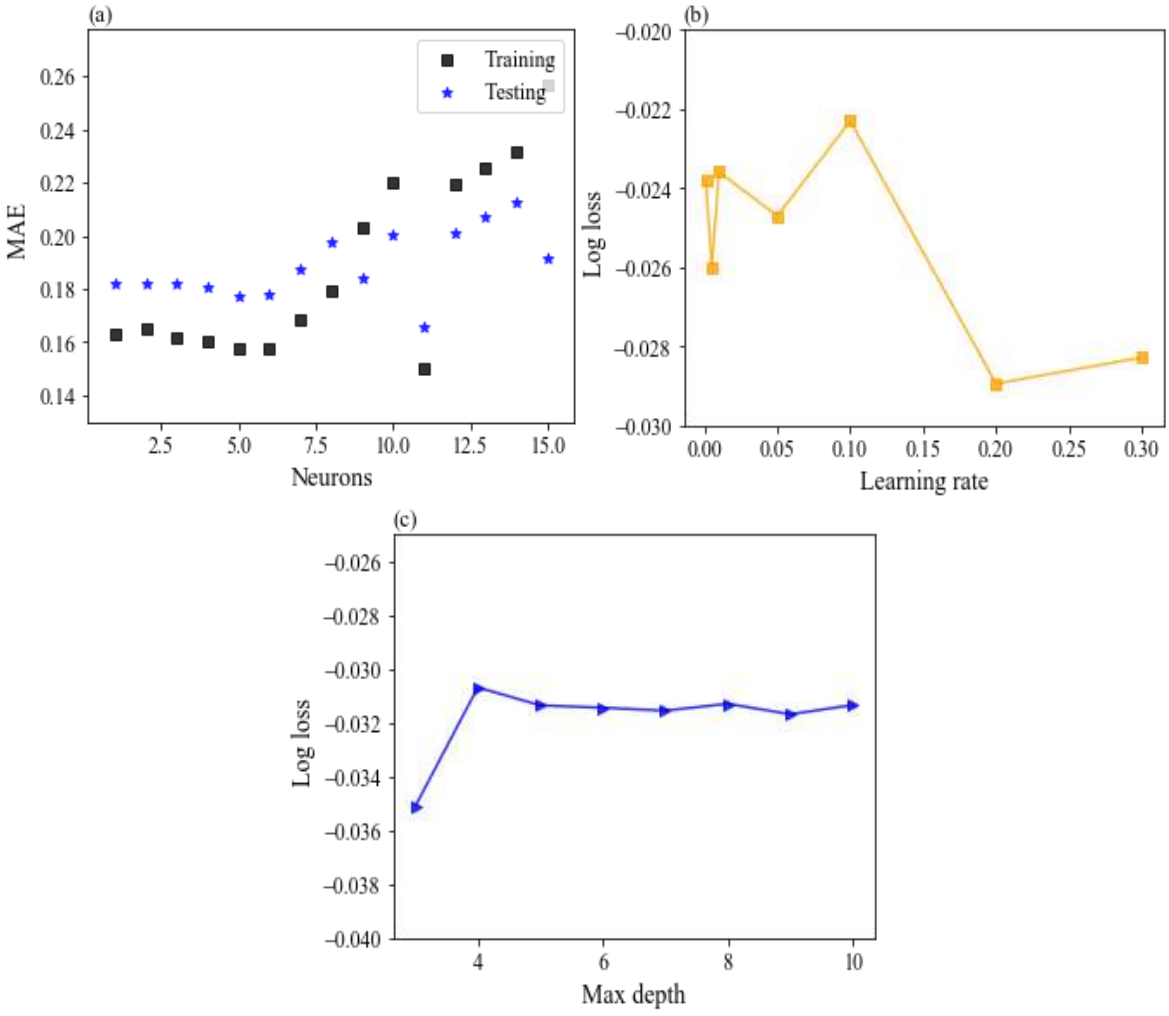


Figure 5. 3. (a) Number of neurons, (b) learning rates (c) max depth, and (d) k values in KNN

Table 5. 2. 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

SDG: stochastic gradient descent, Adam: adaptive momentum, ReLu: Rectified linear unit; a: trees, b: randomised state, c: C, d: ε, e: γ, f: estimators, RBF: radial basis function, Poly: polynomial, g: max depth and learning rate of 4 and 0.1, respectively,

5.5 ML models performance

ANN is regarded as the most extensively used approach for non-trivial problems because of deep learning . However, XGBoost, RFR, and SVR fared better in this investigation. In Figure 5.4a and Figure 5.4b, 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 Figure 5.4c and Figure 5.4d, respectively had a good performance with R^2 in the range of 0.67 - 0.75. MLR in Figure 5.4f 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 Figure 5.5 were used to provide a visualisation of the data distribution and validate these findings. According to an analysis of the distribution shapes in Figure 5.5, 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 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 Figure 5.4 and Figure 5.5 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., 2019a). 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).

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., 2015a).

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., 2014a). 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., 2017b).

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).

5.5.1 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 Figure 5.6. According to the H_0 stated in Equation 3.40, 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 Figure 5.6 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.

5.5.2 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 for this work 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 B.2 and Figure 5.7, respectively. In Figure 5.7, 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.

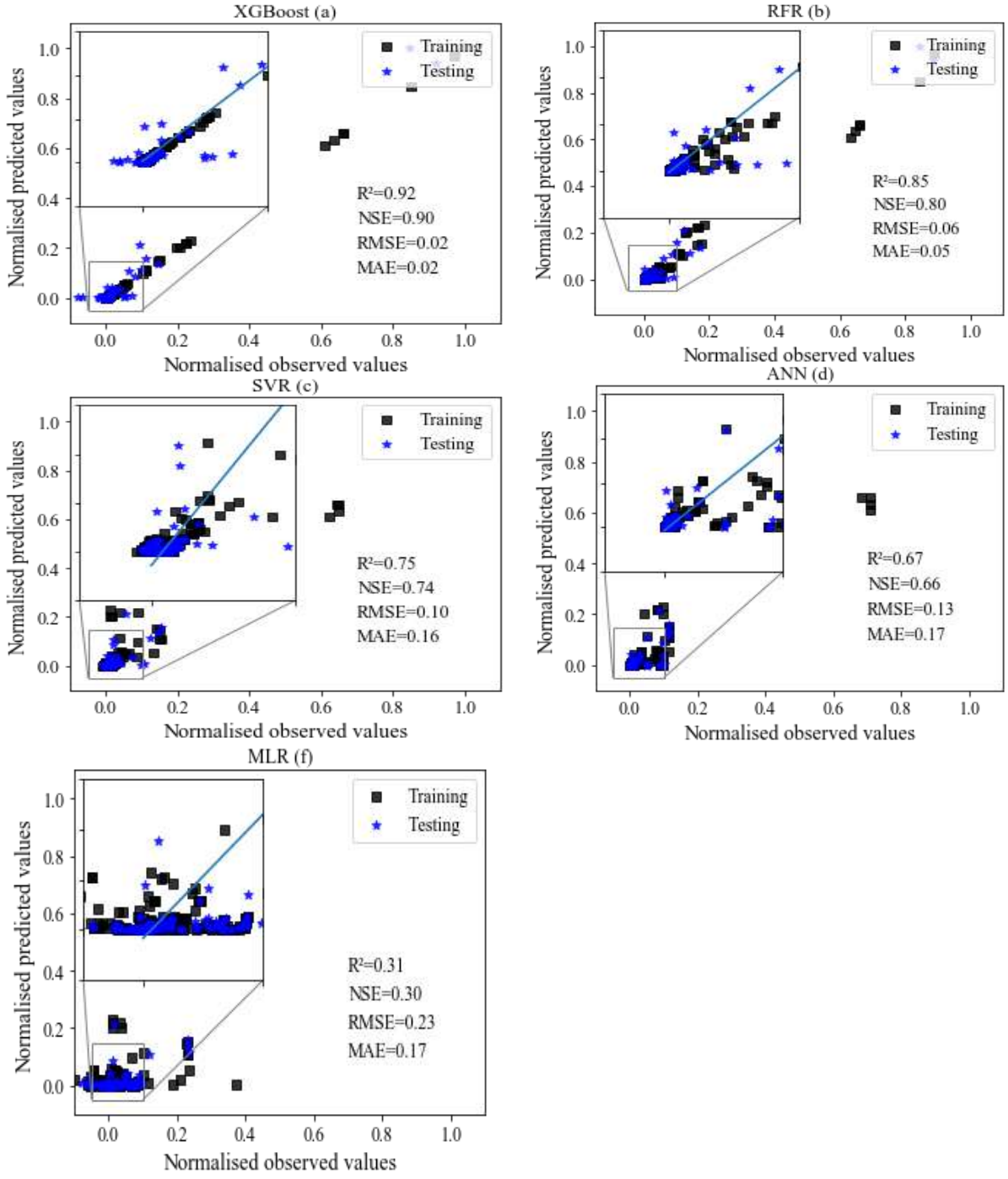


Figure 5. 4. Scatter plots for the predicted models derived for the dissolution of the nZnO data using NOM, time, nZnO concentration, size, IS and pH to the concentration of Zn2+. (a) XGBoost, (b) RFR, (c) SVR, (d) ANN, and (e) MLR.

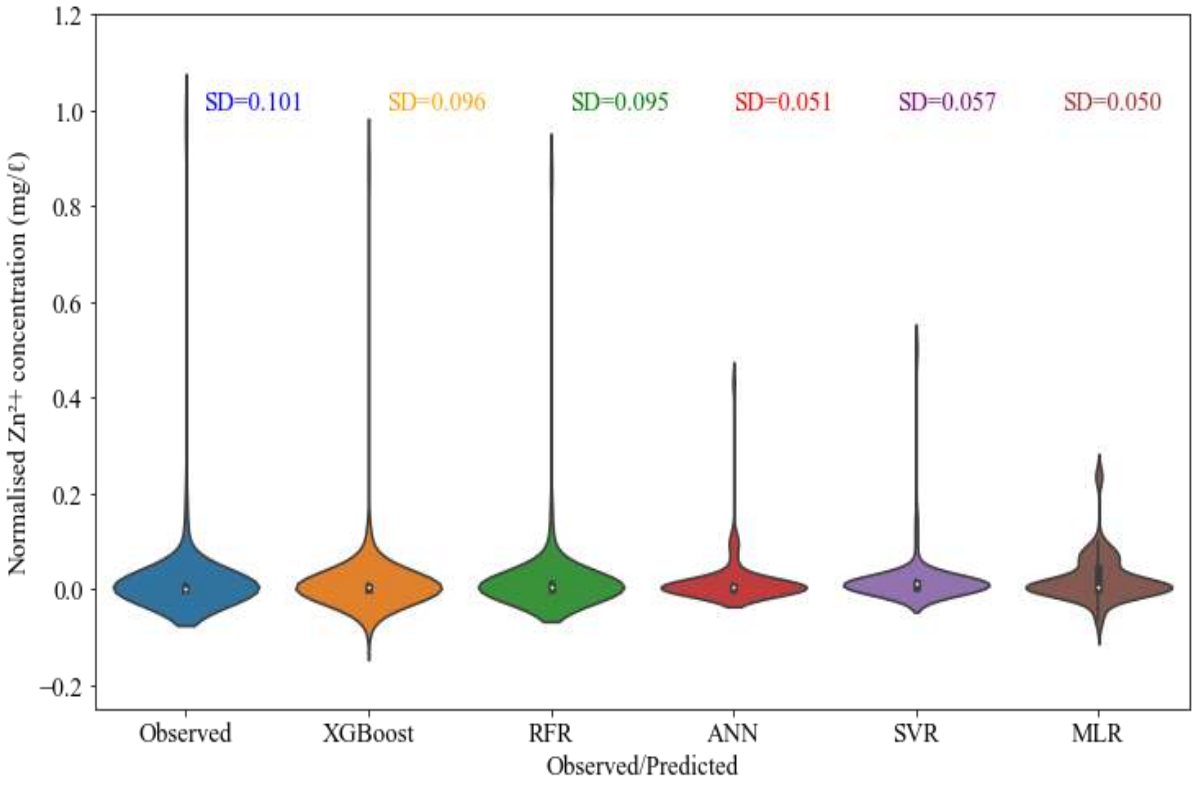


Figure 5. 5. Visualisation of density mass distribution of the predicted values compared to the observed values based on violin plots (VP) (n= 237).

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

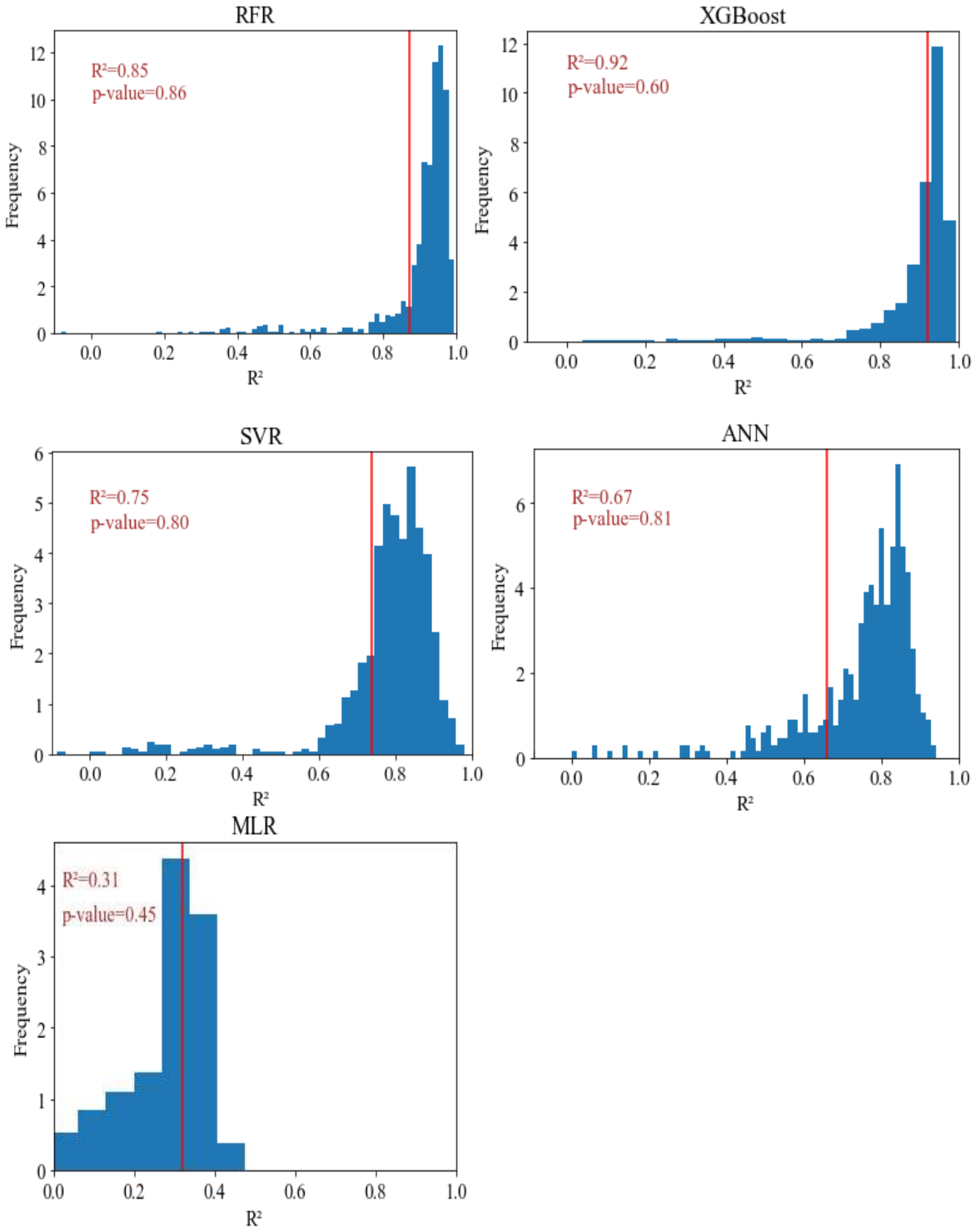


Figure 5. 6. Results of the randomisation test showing the distribution of permutated results (1000 iterations) against the sampled distribution. R^2 (red line) was used as a test statistic.

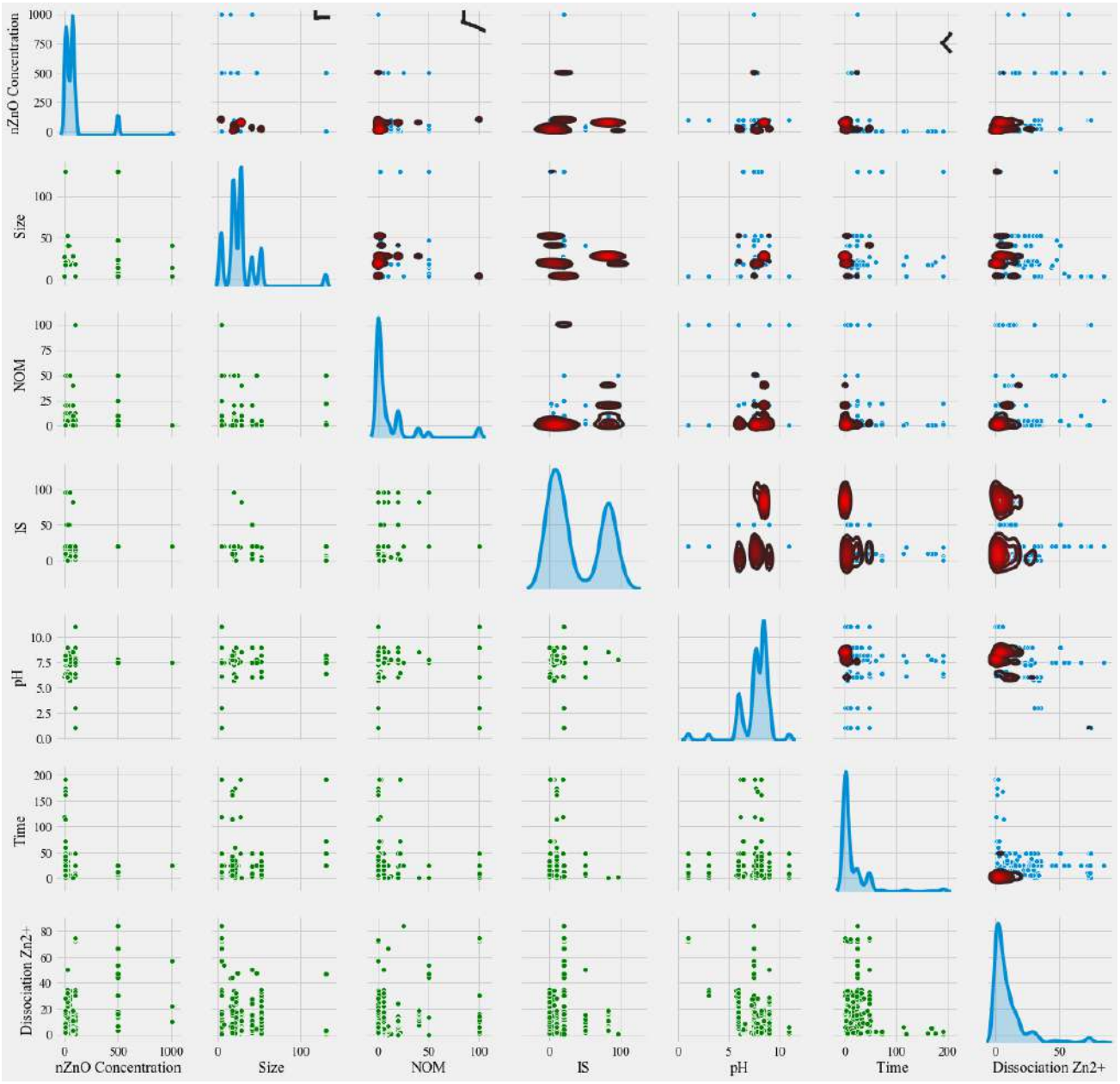


Figure 5. 7. Pair plots showing the density distribution of input and output parameters in training data to characterised AD. The red circle shows a higher distribution

5.6 Chapter summary

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, guides 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 the aqueous environment of among the developed ML models, XGBoost and RFR algorithms were found to be the most effective ML techniques.

Chapter 6. A model for screening the fate and behaviour of the ENPs in aquatic systems using semi-quantitative analysis and decision tree classifiers

This chapter describes the results of SQA integrated with a DT for screening the fate and behaviour of ENPs in aquatic systems. Three specific objectives were pursued. First, a parsimonious hierarchical framework was developed to map multiple input variables to intermediate as well as exposure outputs. Second, using Saaty's rating system the weights were assigned to develop the DT classifiers. Finally, the case studies of nZnO, nAg, and $nTiO_2$ were used to demonstrate the proposed functionality of the framework model, since it is not feasible to cover all ENPs.

6.1 Introduction

ENPs undergo simultaneous chemical and physical transformation mechanisms in the aqueous environments such as the aggregation state (Wagner et al., 2014), dissolution state (Odzak et al., 2017a), and stabilisation state (Danielsson et al., 2018; Philippe and Schaumann, 2014), among others. The behaviour of ENPs in the aqueous environment has a profound effect on their bioavailability, persistence, bioaccumulation, and possible deleterious effects (Leareng et al., 2020; Mahaye et al., 2017). This chapter discusses the results of the use of the SQA integrated with DT for the evaluation of the fate and behaviour in the aqueous environment.

The semi-quantitative models (SQM) are rooted in the use of ranking and scoring factors, rather than the application of mathematical terms used in most mechanistic models (Giubilato et al., 2014; Narita et al., 2014). SQA incorporates less potential uncertainty than quantitative methods or basic qualitative procedures, making it an ideal option in domains with sufficient or poor quantitative data (Obiedat and Samarasinghe, 2016; Tang et al., 2019). The approach does not quantify the exact amount of a given item, a principle similar to qualitative analyses (Amirshenava and Osanloo, 2019; Grella et al., 2019). It uses numerical weights or scores to express the influence on different outcomes based on expert judgment, and intuition rather than the if-then linguistic rules (Singer et al., 2017). On the other hand, DTs are ML

techniques that estimate a target variable's value through the acquisition of basic decision rules or nested if and then rules (Labouta et al., 2019). DT is an effective nonparametric supervised learning method applied to regression and classification problems (Dong et al., 2022). Compared to other algorithms, and more adaptable due to their hierarchical structure, Boolean logic, and representations (Sizochenko et al., 2019). The use of SQA integrated with DTs can be highly valuable for simplifications of ENP exposure.

6.2 Hierarchical framework

A conceptual framework is a hierarchical process that maps significant multiple variables to the targeted output(s) (Saaty, 1987; Van den Brink et al., 2019). Here, a hierarchical framework was applied for two-fold reasons. It is easy to update as new information is generated without the need to completely reconfigure the model. Second, the approach is transparent, and therefore easy to use by non-experts for effective decision-making. The hierarchical model begins with the input, to intermediate variables, and ends with the output, where the decision is to be made (Saaty, 2016). As a common practice, the hierarchical framework should be compact and well-organized to capture the status quo, but small and organised to easily integrate changes (Saaty, 1987).

To derive the number of parameters and establish relationships between various parameters, this work used both the qualitative evidence-based procedure (Tolaymat et al., 2015) and Occam's Razor parsimonious concept (Blumer et al., 1987) described in Section 3.3. The degree of aggregation state ($\alpha_{aggregation}$), colloidal stabilisation ($\alpha_{stabilisation}$), and dissolution state ($\alpha_{dissolution}$) are largely driven by WC and PC, among others. As results at level I of Figure 6. 1 inputs broadly consist of PC (e.g., ZP, size, coating, etc.), and WC properties (e.g., pH, IS, NOM, etc.). Additionally, the solubility and exposure duration parameters were incorporated into the model to take into consideration the diverse ENPs and the dynamics that influence their transformation processes. The impact of these parameters on various ENP transformation processes in freshwater media is extensively covered in numerous review articles (Amde et al., 2017; Baun et al., 2017; Hedberg et al., 2019; Louie et al., 2016; Lowry et al., 2012; Peng et al., 2017; Philippe and Schaumann, 2014; Wagner et al., 2014).

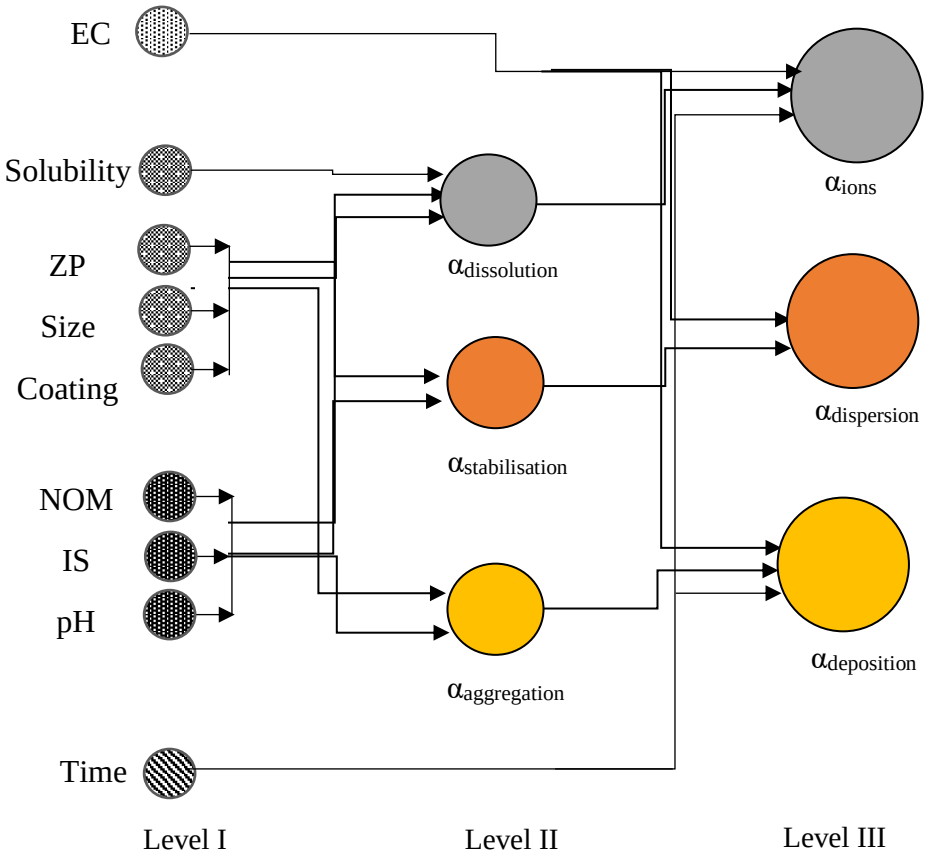


Figure 6. 1. A conceptual structure of the model inputs, intermediate, and output parameters mapping exposure assessment of ENPs in the environment

Furthermore, the exposure potency of ENPs in aqueous media such as bioaccumulation, bioduration, bioavailability, and eventual interactions with biological life-forms in the ecosystems are highly dependent on their degree of dispersion ($\alpha_{dispersion}$), degree of free ions ($\alpha_{dispersion}$) and degree of deposition ($\alpha_{dispersion}$) within a given media (Hamilton et al., 2019). Deposition in aqueous media is a function of the density of the media, the dimension of the fractal, and the weight of the agglomerates (Hartmann et al., 2014). The degree of homo- and hetero-aggregation and attachment efficiency influence ENP deposition (Schaumann et al., 2015).

6.3 Rating of parameters

The identified input parameters in Figure 6.1, were assigned weights to reflect the relative strength that each parameter exerts toward different processes based on

Saaty's rating system range of 1 to 9 (Saaty, 2016). According to the Saaty scale, a higher weight indicates a greater impact of a particular element on the output in question, while a lower weight suggests the contrary (Saaty, 2016). The PEC and analytically MEC are generally estimated to be lower ng/ℓ or µg/ℓ (Zhao et al., 2021). For example, the PEC of $nTiO_2$ in surface water was 0-30 ng/ℓ in the Rhône River, France (Sani-Kast et al., 2015). The nAg in surface water in Europe was 0.87-7.84 ng/ℓ (Sun et al., 2016) and 0.03-2.79 ng/ℓ in freshwater (Giese et al., 2018). About 0.01 to 0.150 µg/ℓ of nZnO concentration was estimated in surface waters (Dumont et al., 2015; Gottschalk et al., 2013). Furthermore, in Taihu Lake in China, the concentrations of nAg and $nTiO_2$ were 0.77 ± 0.24 ng/ℓ and 3.83 ± 0.91 µg/ℓ, respectively (Xiao et al., 2019).

Qualitative levels to describe EC were low (≤ 500 ng/ℓ), moderate (> 500 ≤ 1500 ng/ℓ), and high (> 1500 ng/ℓ) and assigned weights of 1, 3, and 5, respectively (Ramirez et al., 2022). Furthermore, IS and NOM in natural water bodies differ greatly depending on the biogeochemical region (Wagner et al., 2014). NOM and IS in freshwater systems are found in the range between 0 and 30 mg/ℓ (Abbas et al., 2020; Philippe and Schaumann, 2014) and 0 to 10 mM, respectively (Troester et al., 2016). Given that macromolecules are known to be important in the various transformations of ENPs in aqueous media, weights ranging from 1 to 9 were assigned to NOM using the Saaty formalism. IS can compress the electric double layer (EDL) and Debye length through the charge screening effect and consequently, improve the aggregation of ENPs in aqueous media (Peng et al., 2017). Weights in the range of 1 to 7 were assigned to IS. Furthermore, pH ranges from 6.5 to 8.5 in freshwater systems (Troester et al., 2016). The impact of the pH of the aqueous solution was measured by the absolute distance relative to the point of zero charge (PZC). The pH was classified into three categories: high (pH >> PZC << pH), moderate (pH > PZC < pH), and low (PZC ~ pH), and these were assigned weights between 1 and 5.

Table 6. 1. The ranking formalism for exposure model input parameters used to evaluate various intermediate processes

Inputs	Units	Weights	Qualitative values	Ranges
EC	ng/ℓ	1-5	*Low*	≤ 500

			Moderate	> 500 ≤ 1500
			High	> 1500
NOM	mg/l	1-9	*Very low/low*	≤ 3
			Moderate	> 3 ≤ 7
			High	> 7
pH	-	1-5	*low*	≤ 2
			Moderate	>2 ≤ 5
			High	> 5
IS	mM	1-7	*Very Low/Low*	≤ 3
			Moderate	> 3 ≤7
			High	> 7
Size	nm	1-3	*Small*	≤ 30
			Moderate	> 30 ≤ 60
			Large	> 60
ZP	mV	1-3	*Low*	≤ 10
			Moderate	> 10 ≤ 20
			Large	> 20
Coating	-	1-9	*NC*	
			EC	
			SC	
Solubility	%	1-5	*Low*	*≤10*
			Moderate	*<10 ≤ 70*
			High	> 70
Time	h	1-5	*Short*	*≤ 24*
			Moderate	*> 24 ≤ 72*
			Long	*> 72*

The coating materials and the solubility of the ENPs are key PC properties, given their impact on the stability of the ENPs (Alkilany et al., 2016; Louie et al., 2016). Coating agents are generally categorised as electrostatic and steric stabilisers depending on the mode of stabilisation (Lodeiro et al., 2016). Thus, weights between 1 and 9 were

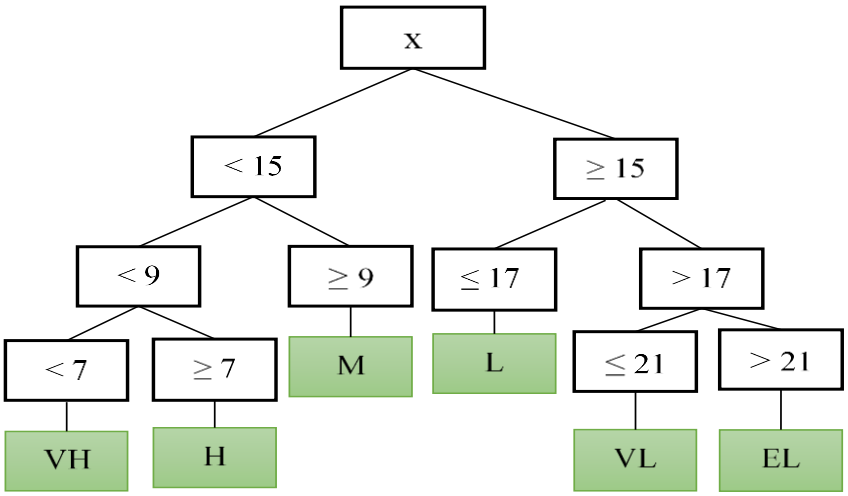
assigned to different coating agents. Moreover, due to the paucity of information about a given ENPs for solubility. Elsewhere, the DF4nanoGrouping initiative suggested that ENPs should be classified as soluble or bio-persistent (Arts et al., 2015). In this study, the solubility was classified into the following qualitative descriptions; high (> 70%), moderate (10 –70%), and low (< 10%), and given weights ranging from 1 to 5.

ENPs are materials with peripheral dimensions in the range of 1- 100 nm. ENPs with a ZP of ± 30 mV are considered stable in aquatic systems (Lowry et al., 2016). The ZP was defined as absolute values using qualitative terms of low (ZP ~ 0), moderate (ZP > 0 < ZP), and high (ZP >> 0 << ZP). In the range of 1-3, the primary size and the ZP received the lowest weights. This was because, in addition, smaller ENPs increase active sites associated with their large surface area (Bian et al., 2011; David et al., 2012). The primary size of the ENPs as specified by the vendors differs considerably from the measured values in the laboratory. Furthermore, for chronic and acute toxicity assessments, exposure times of 96 and 72 hours, respectively, are generally used following the Organisation for Economic Co-operation and Development (OECD) (Macko et al., 2021). Weights ranging from 1 to 5 were assigned to the qualitative levels of the exposure durations.

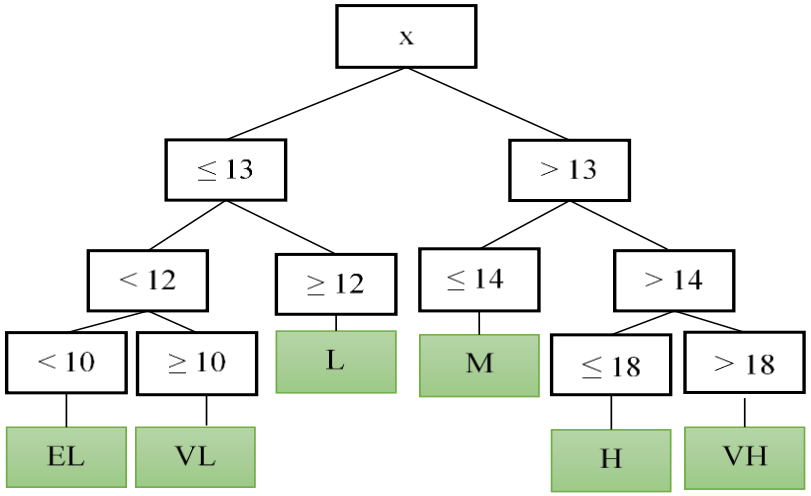
6.4 Decision Tree Classifiers

In this work, the scores defined in the leaf nodes of the decision tree classifiers (DTC) in Figure 6. 2 and Figure 6. 3 were obtained using Equations 3.42-3.44. The evaluation of the model followed pseudocode in Algorithm 11. In Figure 6. 2a, the leaf nodes of less (< 7) and greater (> 21) denote a *very high* and *extremely low* $\alpha_{aggregation}$, respectively. In contrast, Figure 6. 2b indicated that $\alpha_{stabilisation}$ was *extremely low* for a score of less (< 10) and *extremely high* for greater (> 18). The scoring system took into account the permissible environmental outcomes based on an expert understanding of the interactions between input and output. For example, it is impermissible that $\alpha_{stabilisation}$, $\alpha_{aggregation}$, or $\alpha_{dissolution}$ of ENPs can be high simultaneously.

(a)



(b)



(c)

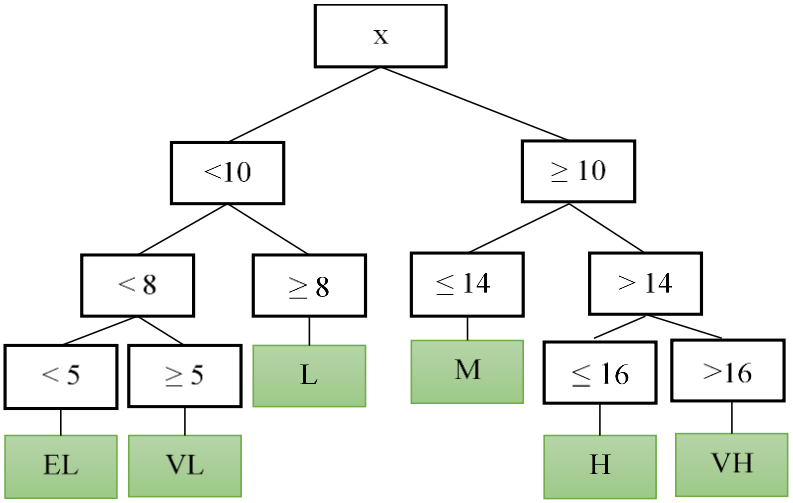


Figure 6. 2. Decision tree for scoring formalism for (a) $\alpha_{aggregation}$, (b) $\alpha_{stabilisation}$ and (c) $\alpha_{dissolution}$

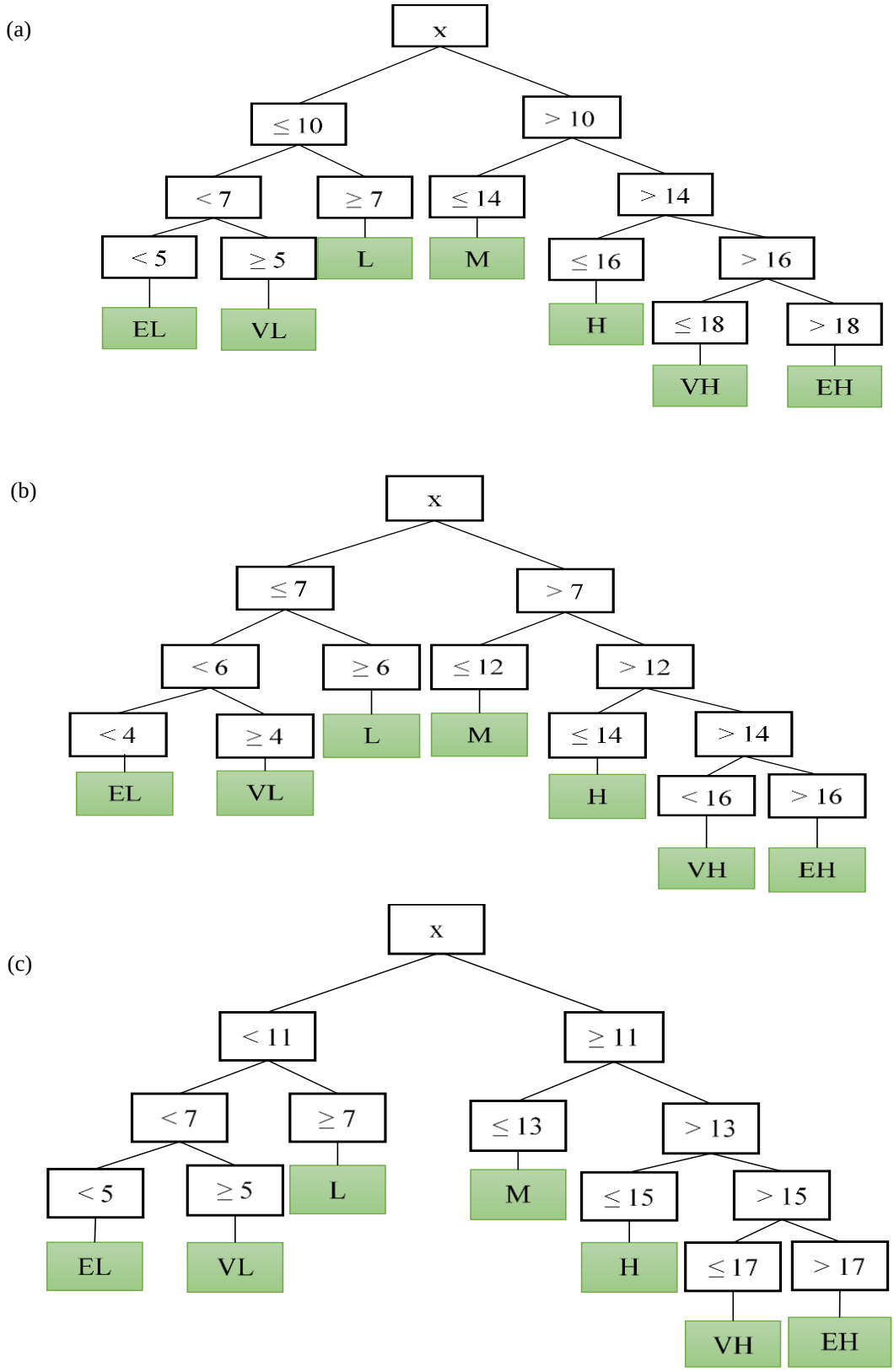


Figure 6. 3. Decision tree for scoring formalism for (a) $\alpha_{deposition}$ (b) $\alpha_{dispersion}$ and (c) $\alpha_{ionic\ species}$ as exposure model outputs.

Furthermore, weights were assigned to the qualitative outputs at level II to generate the sub-criteria for level III. For example, *very high* and *extremely low* $\alpha_{aggregation}$ were assigned weights of 9 and 1, respectively. For rating the $\alpha_{dissolution}$ and $\alpha_{stabilisation}$, the opposite was true, as shown in Table C.1. The leaf nodes of less (< 5) and greater (> 18) in Figure 6. 3a denote *extremely low* and *-high* $\alpha_{deposition}$, respectively. A *very high/high* α_{ions} of ENPs in aqueous media yields low $\alpha_{dispersion}$ and $\alpha_{aggregation}$ (Hedberg et al., 2019). In Figure 6. 3b, the leaf nodes of less (< 4) and greater (> 16) denoted *extremely low* and *-high* $\alpha_{dispersion}$, respectively. Additionally, in Figure 6. 3c, the leaf nodes of less (< 5) and greater (> 17) represented *extremely low* and *-high* α_{ions}, respectively.

6.5 Evaluation of the developed model

6.5.1 Case studies of nAg, nZnO, and $nTiO_2$

To illustrate the functionality of the developed model, case studies of soluble (nAg and nZnO) and non-soluble ($nTiO_2$) were considered. Table 6.2 (nAg), Table 6.3 (nZnO), and Table 6.4 ($nTiO_2$) summarise the inputs to evaluate the developed models. All the conditions applied in Examples I-XII of Tables 6.2 (nAg), were also made to be valid for Tables 6.3 (nZnO), and Table 6.4 ($nTiO_2$). The WC parameters (e.g., NOM, pH, and IS) resemble values encountered in natural systems to show the sensitivity of the created model. For instance, Examples I-XII of Table 6.2 (nAg), Table 6.3 (nZnO), and Table 6.4 ($nTiO_2$) assume a pH of 6.5 and Examples XIII-XIV were based on alkaline pH values of 8.5, respectively. Citrate (CIT) and polyvinylpyrrolidone (PVP) were utilised to represent both electrostatic and steric stabilizers, respectively (Louie et al., 2016; Sharma et al., 2014). Moreover, to reduce the complexity of involving many factors, both absolute ZP and size were assigned constant values for all the scenarios developed.

6.5.2 Results and Discussion

Table 6.5 (nAg), Table 6.6 (nZnO) and Table 6.7 ($nTiO_2$) depict the results using the inputs provided in Table 6.2 (nAg), Table 6.3 (nZnO), and Table 6.4 ($nTiO_2$), respectively. The proposed system was argued and compared against the findings available in literature scientific papers. In Example I, the ECs for nAg, $nTiO_2$ and nZnO,

were 6.3, 192 ng/l, and 84 ng/l, respectively. These ECs were rated as *low* and assigned weights of 1 based on Table 6.2. Further, concerning behaviour the model estimated the $\alpha_{aggregation}$ in Table 6.5 (nAg) and Table 6.6 (nZnO) as *very low*, and the $\alpha_{dissolution}$ and $\alpha_{stabilisation}$ as *high* and *low*, respectively. On the contrary, the $\alpha_{aggregation}$ was qualitatively rated *high* in Table 6.7 ($nTiO_2$). The low colloidal stability predicted for $nTiO_2$ was attributed to the exposure medium pH of 6.5 which falls within the PZC regime typically found between 5.2 and 7 (Loosli et al., 2013). On the other hand, the enhanced $\alpha_{dissolution}$ for nAg and nZnO was attributed to oxidation, which has previously been shown to occur around pH values of 5 and 6 for these ENPs (Han et al., 2014; Peretyazhko et al., 2014). Chen et al. (2016), for example, showed the dissolution of nZnO increases at acidic pH values (< 6.5) and decreases at pH > 7. Since, the PZC is generally found in the range of 8 - 9.5 (Bian et al., 2011) and 2.5 – 4 (Fernando and Zhou, 2019; Sharma et al., 2014) for ZnO and nAg, respectively.

Moreover, the α_{ions} was evaluated to be *moderate* after integrating the ECs and exposure time, which were qualitatively classified as *low* and *short*, respectively, in Table 6.5 (nAg) and Table 6.6 (nZnO). This was the result of the dependence of free ionic species on concentration (Leareng et al., 2020; Musee et al., 2014). Similarly, while $\alpha_{aggregation}$ was assessed as the dominant process in Table 6.7 ($nTiO_2$), the $\alpha_{deposition}$ was estimated as *low.* The collision frequencies between particles depend on the initial concentrations based on Coulomb theory (Hartmann et al., 2014) and deposition can be expected to be low under a short period. Therefore, based on the inputs provided in Example I both nAg and nZnO are likely to dissociate which can increase toxicity in suspension-based organisms such as *algae*. On the other hand, $nTiO_2$ is likely to be immobilised, which will decrease the bioavailability of ENPs in suspension and their harmful effects on water or suspension-based microorganisms (Abbas et al., 2020; Wagner et al., 2014).

Predicted results showed a higher $\alpha_{stabilisation}$ and both the $\alpha_{dissolution}$ and $\alpha_{aggregation}$ were reduced in Examples II-III of Tables 6.5-6.7 as compared to non-coated ENPs in Example I under the same pH of 6.5, IS of 2.4 mM, and NOM of 2.5 mg/l values. However, the CIT-coated ENPs demonstrated reduced stability over that of PVP-coated ENPs. The reduction in the $\alpha_{aggregation}$ and $\alpha_{dissolution}$ was attributed to the presence of high weights assigned to coating agents such as CIT and PVP. Surface

coatings improve the stability of ENPs, and suppress the $\alpha_{aggregation}$ and/or $\alpha_{dissolution}$ in turn, toxicity (Ellis et al., 2016). These findings are consistent with earlier research investigating the impact of PVP and CIT (Ellis et al., 2016; Tejamaya et al., 2012). The higher $\alpha_{stabilisation}$ for PVP-coated ENPs compared to CIT-coated ENPs was due to the high molecular weight of the former, resulting in a strong shielding effect (Li et al., 2012).

In Examples IV-VI of Tables 6.2-6.4, the different ECs were considered; however, they had a relatively insignificant effect on the qualitative rating in Tables 6.5-6.7. Further, with respect to behaviour the results showed a decrease in $\alpha_{stabilisation}$ for non- and CIT-coated ENPs in Examples IV-V of Tables 6.5-6.7. The observed results were attributed to a 4.2-fold increase in IS relative to Examples I-III. Reports by Huynh and Chen. (2011) and Ellis et al. (2016) showed that citrate-nAg exhibited instability under high IS (Ca^{2+} concentrations greater than 9 mmol/L) resulting in hetero aggregation through complexation or bridging. On the other hand, the PVP-coated ENPs in Examples VI of Tables 6.5-6.7 did not show a significant effect with a 4.2-fold increase in IS. This was consistent with previous studies reports, which showed that increasing IS in PVP-coated nAg did not significantly alter hydrodynamic diameter (HDD) due to high molecular weight and shilding effect (Ellis et al., 2016; Tejamaya et al., 2012).

The qualitative ranking for α_{ions} increased for both non-coated and CIT-coated ENPs of nAg, and nZnO, due to the 16-fold increase in duration in Examples VII-VIII and X-XI of Table 6.5 (nAg) and Table 6.6 (nZnO). Similarly, an increase in $\alpha_{deposition}$ was noted in Examples VII-VIII and X-XI of Table 6.7 ($nTiO_2$). This was because long exposure times can promote the hydrolysis of nZnO and improve the sedimentation process (Bhuvaneshwari et al., 2016). Liu et al. (2019), for example, demonstrated that citrate-coated nAg increased dissolution after 96 h of exposure and deleterious effects on zebrafish (*Danio rerio*). On the other hand, in Examples IX and XII of Tables 6.5-6.7 the PVP-coated ENPs did not exhibit discernible alteration with a 16-fold increase in exposure time. These results are in agreement with earlier research by Ellis et al. (2016) and Sharma et al. (2014), which found that PVP-nAg remained stable for 96 hours as the results of a high shielding effect despite citrate-nAg displaying significant colloidal instability.

Table 6. 2. Set of model inputs data randomly sourced from published literature to formulate scenarios for nAg.

No.s	River and country	EC (ng/l)	Coating	Size (nm)	ZP (mV)	NOM (mg/l)	PZC	pH -	pH	IS (Ca^{2+}, Mg^{2+}) (mM	Time (h)	Sol %
I	(Meuse Holland	6.3*)[a]	NC	30	-20	2.5	3	6.5	3.5	2.4	6	80
II	(Meuse Holland	6.3*)[a]	CIT	30	-20	2.5	3	6.5	3.5	2.4	6	80
III	(Meuse Holland	6.3*)[a]	PVP	30	-20	2.5	3	6.5	3.5	2.4	6	80
IV	(IJssel Holland	2.2)[a]	NC	30	-20	2.5	3	6.5	3.5	10	6	80
V	(IJssel Holland	2.2)[a]	CIT	30	-20	2.5	3	6.5	3.5	10	6	80
VI	(IJssel Holland	2.2)[a]	PVP	30	-20	2.5	3	6.5	3.5	10	6	80
VII	(Pre-alpine lakes Germany	2.35**)[b]	NC	30	-20	2.5	3	6.5	3.5	2.4	96	80
VIII	(Pre-alpine lakes Germany	2.35**)[b]	CIT	30	-20	2.5	3	6.5	3.5	2.4	96	80
IX	(Pre-alpine lakes Germany	2.35**)[b]	PVP	30	-20	2.5	3	6.5	3.5	2.4	96	80
X	(Pre-alpine lakes Germany	2.35**)[b]	NC	30	-20	2.5	3	6.5	3.5	10	96	80
XI	(Pre-alpine lakes Germany	2.35**)[b]	CIT	30	-20	2.5	3	6.5	3.5	10	96	80
XII	(Pre-alpine lakes Germany	2.35**)[b]	PVP	30	-20	2.5	3	6.5	3.5	10	96	80
XIII	(Taihu Lake China	0.77)[d]	NC	30	-20	2.5	3	8.5	5.5	2.4	72	80
XIV	(Taihu Lake China	0.77)[d]	NC	30	-20	10.0	3	8.5	5.5	2.4	72	80

pH = abs ($pH_{measured}$ - $PZC_{medium\ value}$), ZP = abs ($ZP_{measured}$). a: (Peters et al., 2018), b: (Wimmer et al., 2018), c: (Xiao et al., 2019)

Table 6. 3. Set of model inputs data randomly sourced from published literature to formulate scenarios for nZnO.

No.s	River and country	EC (ng/l)	Coating	Size (nm)	ZP (mV)	NOM (mg/l)	PZC	pH_m -	pH	IS (Ca^{2+}, Mg^{2+}) (mM	Time (h)	Sol %
I	(USA	0.192*)[a]	NC	30	-20	2.5	9	6.5	2.5	2.4	6	80
II	(USA	0.192*)[a]	CIT	30	-20	2.5	9	6.5	2.5	2.4	6	80
III	(USA	0.192*)[a]	PVP	30	-20	2.5	9	6.5	2.5	2.4	6	80
IV	(Switzerland	32**)[b]	NC	30	-20	2.5	9	6.5	2.5	10	6	80
V	(Switzerland	32**)[b]	CIT	30	-20	2.5	9	6.5	2.5	10	6	80
VI	(Switzerland	32**)[b]	PVP	30	-20	2.5	9	6.5	2.5	10	6	80
VII	(Europe	190**)[c]	NC	30	-20	2.5	9	6.5	2.5	2.4	96	80
VIII	(Europe	190**)[c]	CIT	30	-20	2.5	9	6.5	2.5	2.4	96	80
IX	(Europe	190**)[c]	PVP	30	-20	2.5	9	6.5	2.5	2.4	96	80
X	(Europe	190**)[c]	NC	30	-20	2.5	9	6.5	2.5	10	96	80
XI	(Europe	190**)[c]	CIT	30	-20	2.5	9	6.5	2.5	10	96	80
XII	(Europe	190**)[c]	PVP	30	-20	2.5	9	6.5	2.5	10	96	80
XIII	(USA	0.192*)[a]	NC	30	-20	2.5	9	8.5	0.5	2.4	72	80
XIV	(USA	0.192*)[a]	NC	30	-20	10.0	9	8.5	0.5	2.4	72	80

MEC* PEC**, a: (J.-S. Choi et al., 2018), b: Gottschalk et al., 2011), c: (Dumont et al., 2015)

Table 6. 4. Set of model inputs data randomly sourced from published literature to formulate scenarios for $nTiO_2$.

No.s	Rivers&Country	EC (mg/l)	Coating	Size (nm)	ZP (mV)	NOM (mg/l)	PZC	pH_m -	pH	IS (Ca^{2+}, Mg^{2+}) (mM)	Time (h)	Sol %
I	(Swimming pool, US	84*)[a]	NC	30	-20	2.5	6	6.5	0.5	2.4	6	10
II	(Swimming pool, US	84*)[a]	CIT	30	-20	2.5	6	6.5	0.5	2.4	6	10
III	(Swimming pool, US	84*)[a]	PVP	30	-20	2.5	6	6.5	0.5	2.4	6	10
IV	(Salt River, US	399*)[a]	NC	30	-20	2.5	6	6.5	0.5	10	6	10
V	(Salt River, US	399*)[a]	CIT	30	-20	2.5	6	6.5	0.5	10	6	10
VI	(Salt River, US	399*)[a]	PVP	30	-20	2.5	6	6.5	0.5	10	6	10
VII	(Rhône River, France	30**)[b]	NC	30	-20	2.5	6	6.5	0.5	2.4	96	10
VIII	(Rhône River, France	30**)[b]	CIT	30	-20	2.5	6	6.5	0.5	2.4	96	10
IX	(Rhône River, France	30**)[b]	PVP	30	-20	2.5	6	6.5	0.5	2.4	96	10
X	(Johannesburg City, South Africa	267.3**)[c]	NC	30	-20	2.5	6	6.5	0.5	10	96	10
XI	(Johannesburg City, South Africa	267.3**)[c]	CIT	30	-20	2.5	6	6.5	0.5	10	96	10
XII	(Johannesburg City, South Africa	267.3**)[c]	PVP	30	-20	2.5	6	6.5	0.5	10	96	10
XIII	(Freshwater, Denmark	99.4)[e]	NC	30	-20	2.5	6	8.5	2.5	2.4	72	10
XIV	(Freshwater, Denmark	99.4)[e]	NC	30	-20	10.0	6	8.5	2.5	2.4	72	10

MEC* PEC** a: , b: (Sani-Kast et al., 2015), c: (Musee, 2011) , d: (Gottschalk et al., 2015)

Table 6. 5.A complete set of qualitative rankings for inputs is provided in Table 6.2

ENP type	Ex. No.s	ECs	$\alpha_{aggregation}$	$\alpha_{dissolution}$	$\alpha_{stabilisation}$	$\alpha_{deposition}$	α_{ions}	$\alpha_{dispersion}$
nAg	I	*low*	*Very low*	*High*	*Low*	*Very low*	*Moderate*	*Low*
nAg	II	*low*	*Very low*	*Low*	*High*	*Extreme low*	*Very low*	*Moderate*
nAg	III	*low*	*Extreme low*	*Very low*	*Very high*	*Extreme low*	*Extreme low*	*High*
nAg	IV	*low*	*Moderate*	*Low*	*Extreme low*	*Low*	*Very low*	*Low*
nAg	V	*low*	*Moderate*	*Very low*	*Low*	*Low*	*Extreme low*	*Low*
nAg	VI	*low*	*Low*	*Very low*	*High*	*Very low*	*Extreme low*	*Moderate*
nAg	VII	*low*	*Low*	*High*	*Low*	*Low*	*High*	*Very low*
nAg	VIII	*low*	*Very low*	*Low*	*High*	*Low*	*Moderate*	*Low*
nAg	IX	*low*	*Extreme low*	*Very low*	*Very high*	*Low*	*Low*	*High*
nAg	X	*low*	*Moderate*	*Low*	*Extreme low*	*Moderate*	*Low*	*Extreme low*
nAg	XI	*low*	*Moderate*	*Very low*	*Low*	*Moderate*	*Low*	*Very low*
nAg	XII	*low*	*Low*	*Very low*	*High*	*Low*	*Low*	*Moderate*
nAg	XIII	*low*	*Low*	*Very low*	*High*	*Low*	*Low*	*Moderate*
nAg	XIV	*low*	*Extreme low*	*Very low*	*Very high*	*Very low*	*Very low*	*High*

Legend: Extreme low, Very low, Low, moderate, high, Very low

Table 6. 6. A complete set of qualitative rankings for inputs is provided in Table 6. 3

ENP type	Ex No.s	*ECs*	$\alpha_{aggregation}$	$\alpha_{dissolution}$	$\alpha_{stabilisation}$	$\alpha_{deposition}$	α_{ions}	$\alpha_{dispersion}$
nZnO	I	*Low*	*low*	*High*	*Low*	*Very low*	*Moderate*	*Low*
nZnO	II	*Low*	*Very low*	*Low*	*High*	*Very low*	*Low*	*Moderate*
nZnO	III	*Low*	*Extreme low*	*Very low*	*Very high*	*Extreme low*	*Very low*	*High*
nZnO	IV	*Low*	*Moderate*	*Low*	*Extreme low*	*Low*	*Very low*	*Low*
nZnO	V	*Low*	*Moderate*	*Very low*	*Low*	*Low*	*Low*	*Low*
nZnO	VI	*Low*	*Low*	*Very low*	*High*	*Very low*	*Very low*	*Moderate*
nZnO	VII	*Low*	*Low*	*High*	*Low*	*Low*	*Moderate*	*Very low*
nZnO	VIII	*Low*	*Very low*	*Low*	*High*	*Low*	*Moderate*	*Low*
nZnO	IX	*Low*	*Extreme low*	*Very low*	*Very high*	*Low*	*Low*	*High*
nZnO	X	*Low*	*Moderate*	*Low*	*Extreme low*	*Moderate*	*Low*	*Extreme low*
nZnO	XI	*Low*	*Moderate*	*Very low*	*Low*	*Moderate*	*Low*	*Very low*
nZnO	XII	*Low*	*Low*	*Very low*	*High*	*Low*	*Low*	*Moderate*
nZnO	XIII	*Low*	*High*	*Very low*	*Very low*	*Moderate*	*Low*	*Very low*
nZnO	XIV	*Low*	*Very low*	*Very low*	*Very high*	*Low*	*Very low*	*Moderate*

Table 6. 7. A complete set of qualitative rankings for inputs is provided in Table 6. 4

ENP type	Ex No.s	*ECs*	$\alpha_{aggregation}$	$\alpha_{dissolution}$	$\alpha_{stabilisation}$	$\alpha_{deposition}$	α_{ions}	$\alpha_{dispersion}$
$nTiO_2$	I	*Low*	*High*	*Very low*	*Very low*	*Low*	*Very low*	*Low*
$nTiO_2$	II	*Low*	*Low*	*Extreme low*	*High*	*Very low*	*Extreme low*	*Moderate*
$nTiO_2$	III	*Low*	*Very low*	*Extreme low*	*Very high*	*Very low*	*Extreme low*	*High*
$nTiO_2$	IV	*Low*	*High*	*Extreme low*	*Extreme low*	*Moderate*	*Extreme low*	*Low*
$nTiO_2$	V	*Low*	*Moderate*	*Extreme low*	*Very low*	*Low*	*Extreme low*	*Low*
$nTiO_2$	VI	*Low*	*Low*	*Extreme low*	*High*	*Very low*	*Extreme low*	*Moderate*
$nTiO_2$	VII	*Low*	*High*	*Very low*	*Very low*	*Moderate*	*Very Low*	*Extreme low*
$nTiO_2$	VIII	*Low*	*Low*	*Extreme low*	*High*	*Low*	*Very low*	*Moderate*
$nTiO_2$	IX	*Low*	*Very low*	*Extreme low*	*Very high*	*Low*	*Very low*	*High*
$nTiO_2$	X	*Low*	*High*	*Extreme low*	*Extreme low*	*High*	*Very low*	*Extreme low*
$nTiO_2$	XI	*Low*	*Moderate*	*Extreme low*	*Very low*	*Moderate*	*Very low*	*Extreme low*
$nTiO_2$	XII	*Low*	*Low*	*Extreme low*	*High*	*Low*	*Very low*	*Moderate*
$nTiO_2$	XIII	*Low*	*Low*	*Low*	*High*	*Very low*	*Very low*	*Moderate*
$nTiO_2$	XIV	*Low*	*Low*	*Low*	*Very high*	*Extreme low*	*Extreme low*	*High*

The model predicted *high* $\alpha_{aggregation}$ in Table 6.6 (nZnO) for Example XIII as the pH of 8.5 belongs to the pHpzc regime for nZnO. In contrast, the higher $\alpha_{stabilisation}$ in Table 6.7 was because, at alkaline pH values for $nTiO_2$ has a negative net surface charge resulting in strong repulsion forces, and high particle electrophoretic mobility (M. Zhu et al., 2014a). On the other hand, the higher $\alpha_{stabilisation}$ for nAg in Table 6.5 at alkaline pH values was because the surface of nAg being deprotonated resulting in the formation of an oxide layer (Fernando and Zhou, 2019).

The $\alpha_{stabilisation}$ and $\alpha_{dispersion}$ were rated as *very high* and *high*, respectively in Examples XIV of Table 6.5 and Table 6.7 due to the 4-fold increase in NOM concentration. High concentrations of macromolecules like NOM induce both steric and electrostatic stresses on the surface of ENPs (Philippe and Schaumann, 2014; Wagner et al., 2014). This, in turn, can result in locked oxidation sites by the formation of NOM-coatings on ENP suppressing the dissolution of nAg toxic ions and subsequently their harmful effects (Gunsolus et al., 2015; Kennedy et al., 2010). Higher concentrations of dissolved organic matter (DOM), for example, reduced the toxicity of *Ceriodaphnia dubia* and *Pseudokirchneriella subcapitata* (McLaughlin and Bonzongo, 2012).

6.6 Model generalisation and limitation

The model presented here is meant to serve as a proof of concept regarding the application of an analytical tool for the prioritization of ENPs risks in the environment. The development of this model required only a qualitative understanding of the interactions between ecosystem components (Pilone and Demichela, 2018). The approach is likely to improve many problems associated with qualitative techniques and the lack of homogeneous quantitative data due to the lack of protocols. In addition, the approach can be used retrospectively to decide on likely exposure levels until comprehensive and robust risk assessment methods can be developed as more information on fate, ecotoxicity and environmental concentration becomes available.

However, it is challenging to rigorous validate the modelled values due to the lack of trace analytical techniques tailored to the detection and quantification of ENPs. As a result, the applicability domain or generalisation of the model can be limited to the ranges considered. Thus, we advise that additional data be taken into account and that the model's output be validated by comparing it with real observed data in future work. The strength of the suggested validation approach is that the comparability of

modelling and actual measurement data may address challenges related to the practicalities associated with the implementation of the proposed screening approach.

In addition, taking into account the evolving nature of data and knowledge in exposure assessment of ENPs in aquatic systems especially as more fate studies are published based on actual environmental matrixes (e.g. river water, lake water, etc.); the authors acknowledge that the proposed framework may change but substantial deviations are unlikely if currently accessible data to date are taken as a guide on the subject matter in this domain. Any deviations may likely be due to complexities and high variability of freshwater systems chemistry that can alter the transformation of ENPs in ways not predictable based on synthetic and simplified exposure media mostly found in reviewed articles used in our study.

However, besides the use of weights and scores in SQA was very helpful for the simplification of complex problems; the approach is also characterised by several limitations. The philosophy of qualitative techniques is based on descriptive approaches and on the paucity of standard rules to assign qualitative levels; thus, is subjective to the degree of uncertainty (Amirshenava and Osanloo, 2019; Grella et al., 2019; Pang and Coghill, 2015). Assigning more qualitative levels increases the sensitivity, but could complicate the problem more. In addition, EC in the Salt River and swimming pool in the US, the concentrations are 84 and 399 ng/l, yet they were both rated as *low* and assigned a weight of 1. In future studies, the subjective element of the developed model should be reduced by integrating mathematical principles such as fuzzy theory, which uses membership function values to define different values with a particular and can handle uncertainty and lack of defined or clear boundaries (Zadeh, 1965).

6.7 Environmental Significance and model deployment

ENPs in aquatic systems undergo one of three processes: sedimentation, dissociation, or dispersion. Depending on the trophic level of the organism under study, the ions, particles, and even aggregates may results in harmful to aquatic organisms to a certain degree (Grillo et al., 2015). In this context, understanding the likely behaviour of ENPs in given WC and PC properties is critical to determine their environmental implications (Abbas et al., 2020). Hartmann et al. (2010), for example, noted that increased hetero-aggregation decreased the negative impacts on *Pseudokirchneriella subcapitata*, a

freshwater green algae, and thus decreased the bioavailability of $nTiO_2$. Immobilisation of micrometre-sized agglomerates of Fe_2O_3, $nTiO_2$, and nAg decreases their exposure to aquatic receptors in the water column (Chekli et al., 2015). Agglomeration and deposition processes decreased the accumulation and toxicity of $nTiO_2$ to microalgae such as *Isochrysis galbana* cells (Hu et al., 2018). This was because agglomerated particles are generally less reactive than their parent particles.

The use of SQA integrated with the rule base system in the form of DT reported in this work provides a starting point for the development of an easy-to-use tool that can be very helpful in supporting the cost-effective determination and screening of the fate and behaviour of ENPs. This, in turn, can reduce the costs associated with undertaking experimental tests for each variation of ENP using various aquatic permutations, since it is not feasible to carry out experimental testing because of a wide range of physicochemical parameters, as well as dynamic interactions with natural colloid particles.

The model uses a modular approach and is an easy-to-use tool based on Microsoft Office as an Excel spreadsheet that is fully automated with less computation. The ranges can be varied with ease, for application in different water systems without completely changing the whole model. The tool deals with the knowledge of various complex problems in an intuitive, appropriate, and flexible way that other intelligent systems cannot. It supports to optimise the optimal use of existing both qualitative and quantitative data features to determine the exposure of ENPs to the fate and behaviour of ENPs in aquatic systems as a basis for supporting decision-making in pursuit of the achievement of sustainable use of nanotechnology-based applications and bridging the knowledge gap between experimentalists and modellers.

6.8 **Chapter Summary**

In summary, this work presents a model developed by the integration of SQA with a series of if-then represented using DT to facilitate the systematic screening of ENP exposure potential and prospective effects on various tropic levels in aquatic systems. Using the Saaty ranking approach; the influence of each input weighted demonstrates that the developed model can be highly valuable in serving as a tool for decision-making and policy formulation. Additionally, the authors further acknowledge that the hierarchical framework developed aided to support EA, but, is by no means complete.

Such a model is valuable given the uncertainty, unstructured, data gaps, lack of primary raw data, and insufficient understanding of the number of variables that influence the exposure potency of ENPs, among others. Therefore, other potential transformations e.g. sulfidation for Ag in natural systems or properties such as shape can be easily updated without completely changing the framework in future work.

Chapter 7. A model using fuzzy logic for assessing the fate and behaviour of metal-based engineered nanoparticles in the freshwater environment

This chapter describes the development of a fuzzy decision-making system (FDMS) for the EA of ENPs in aquatic systems. The objectives of this chapter have been addressed in three-fold, namely; (i) by developing the hierarchical framework to highlight factors that drive the fate and behaviour of ENPs in a freshwater aquatic environment, (ii) by developing the IF-THEN linguistic rules to encapsulate the expert knowledge within FIS and (iii) functionality of the FL based system has been demonstrated by predicting the likely behaviour of nZnO and nTiO_2 as case studies in various freshwater aqueous matrices.

7.1 Introduction

Chapters 4 – 6 discussed several approaches proposed to elucidate the transformation processes of ENPs in aquatic environments**.** ML described in **Chapters 4 and 5** tools have been remarkably successful in discovering scientific information and providing a coherent understanding of the mechanisms that drive the transformation of ENPs (Yalezo et al., 2024). However, the use of ML approaches for data mining and/or developing prediction algorithms, so far, has challenges related to substantive dependence on quantitative data and is not flexible to model partial truth information (Basei et al., 2019b; Yalezo and Musee, 2023). As a result, identifying and even providing the accurate measurement of these transformation processes individually and/or in combination remain difficult; but they are likely to improve in the near future with an increase in the development of curated nano databases, such as NanoE-Tox (Juganson et al., 2015), and the S2NANO database (www.s2nano.org) (Trinh et al., 2018).

On the other hand, semi-quantitative based models described in **Chapter 6** can simplify complex problems, however, they are concomitant with over-simplification and do not account for the uncertainty that arises with the modelling of expert intuitions or perceptions using scores or weights (Ye et al., 2020). So, in light of the challenges

witnessed in **Chapters 4 - 6**, which include a lack of quantitative data, poorly defined information, a lack of sharp boundaries and frameworks to address uncertainty, and the multiplicity of intricate components at play. Alternatively, the use of fuzzy logic (FL) can be a suitable modelling approach to screen the exposure of ENPs and support the reasoning of well-thought-out solutions to complex problems (Lu et al., 2022). Therefore this chapter discusses results on the use of the FL reasoning mechanism to integrate uncertainty and simplify the assessment of the fate and behaviour of ENPs in a freshwater aquatic environment.

7.2 Model input, and output parameters

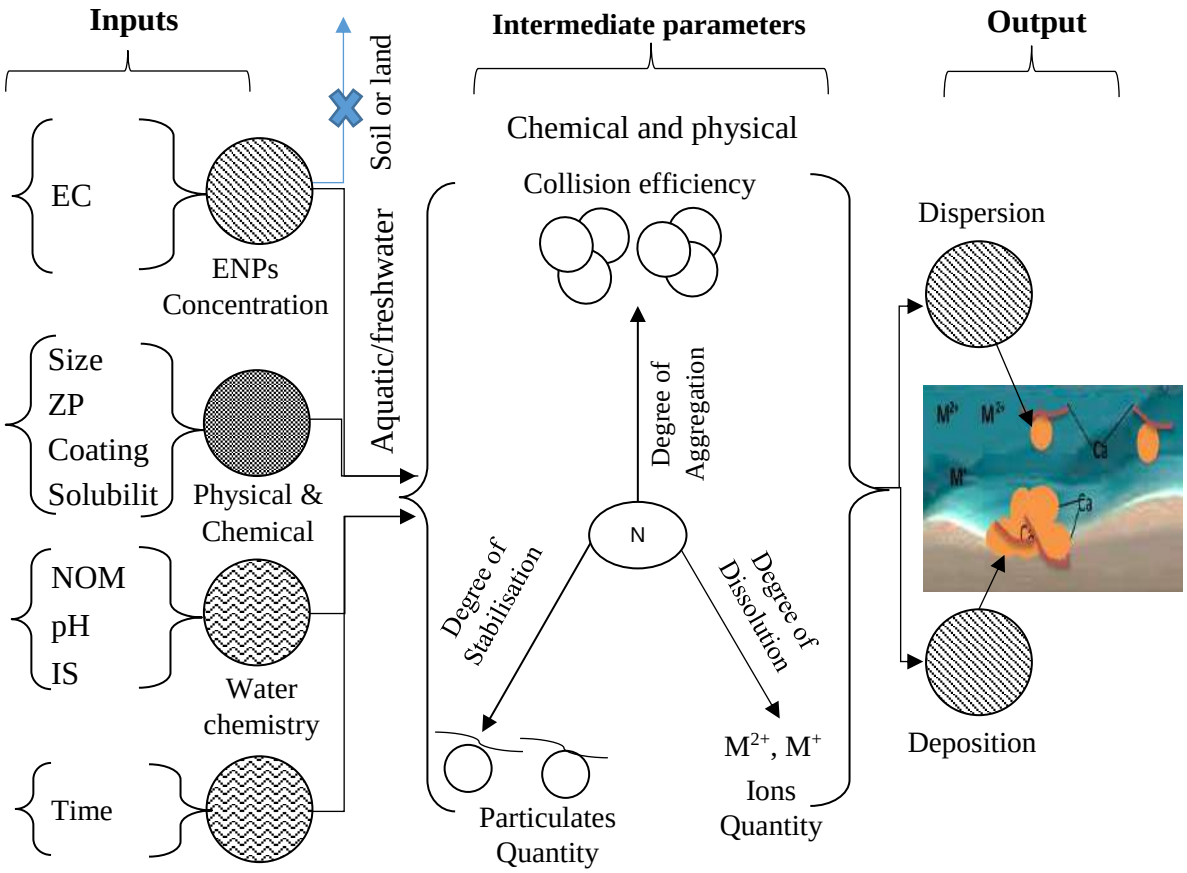


Figure 7. 1. Schematic diagram showing the parameters that influence the exposure of ENPs in aquatic systems.

The fate and transformation behaviour of ENPs in freshwater aquatic environments and their degree of exposure is influenced, among others, by their inherent PC and

receiving WC properties. Figure 7.1 shows the hierarchical framework, which maps the preponderant input variables to intermediates, and target output variables following Occam's razor and evidence-based principle described in Section 3.3. The inputs included the ENP EC (X_1, mg/ℓ), duration (X_2, h), NOM (X_3, mg/ℓ), IS (X_4, mM), size (X_5, nm), pH (X_6, dimensionless), solubility (X_7, %), and ZP (X_8, mv). Various chemical and physical transformation processes, such as the aggregation state (XY_1), the dissolution state (XY_2), and the stabilisation state (XY_3) were assigned as possible intermediate parameters, and the model outputs are dispersion (Y_1) or deposition (Y_2).

7.3 Implementation of FL model

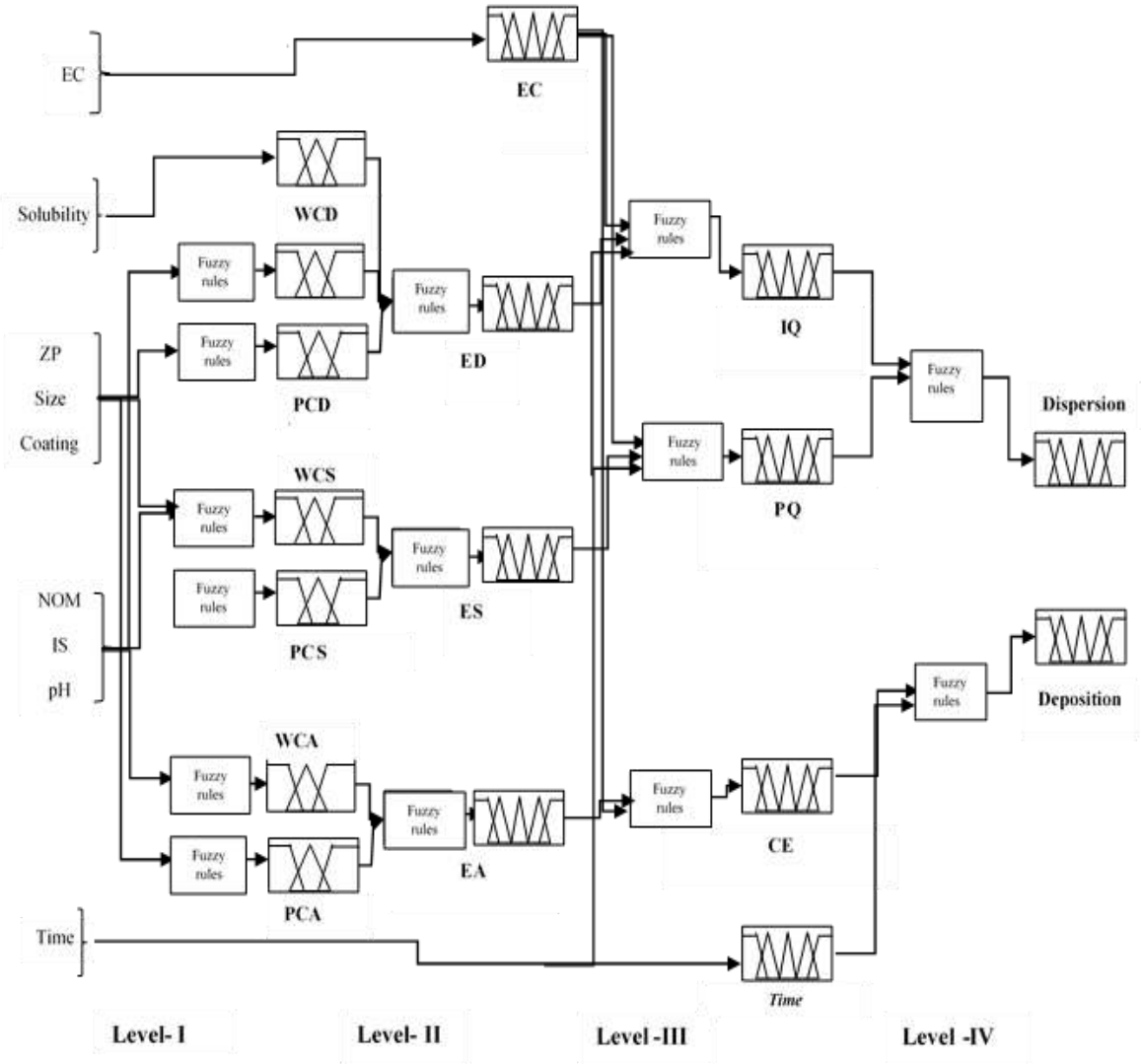


Figure 7. 2. A hierarchical framework for FL model

The Mamdani and Assilian FIS built-in fuzzy logic toolbox integrated into the Matrix Laboratory software (MATLAB, R2007) was used for developing the fuzzy logic models following the *pseudo-code* described in **Algorithm 12**. Implementation and development of fuzzy systems can result in many rules, which can be challenging to manage and develop (Gacto et al., 2011). To reduce the possibility of an unmanageable total number of linguistic rules, the fuzzy hierarchical framework (FHF) described in Figure 7.2 with numerous independent information structures of MISO (Multiple Input, Single Output) was followed. The FHF in Figure 7.2 is comprised of several interconnected fuzzy subsystems designed to use the output(s) of preceding rule-base (s) to serve as input(s) for succeeding ones.

Level I. At the beginning of the FHF, the model had eight deterministic and one qualitative input. Figure 7.3 illustrates the triangular- and trapezoidal-shaped distribution curves for the input parameters. FL is a multivalued model and allows sets between two MFs to overlap. The EC of ENPs in surface waters is generally within the lower band range of µg/ℓ and ng/ℓ (Zhao et al., 2021). In Figure 7.3a, the EC was defined in the range of 0-10000 ng/ℓ or 10 µg/ℓ. A similar formalisation has been shown to be successful in research by Ramirez et al. (2022) using a case study of silver nanoparticles.

In Figure 7.3b and Figure 7.3c, the NOM and IS were defined in the concentration ranges of (≤ 30 mg/ℓ of carbon/ℓ) and ≤ 10 mM, respectively which represent values relevant to freshwater (Abbas et al., 2020; Louie et al., 2016). ENPs with a ZP of ± 30 mV and beyond are considered stable in aquatic systems (Lowry et al., 2016). The pH range is between 6.5 and 8.5 for freshwater systems (Troester et al., 2016). In Figure 7.3d and Figure 7.3e the pH and ZP, respectively were represented as absolute values. In Figure 7.3f, the solubility was classified based on the percentage (0 -100%) of the total released metal concentration. The coating variable in Figure 7.3g was defined in the range of 0 - 1. Additionally, the exposure period in Figure 7.3i was defined in the range of 0-120 hours to be inclusive of 96 and 72 hours, which are generally used to assess acute and chronic toxicity following the Organisation for Economic Co-operation and Development (OECD) guidelines (Macko et al., 2021).

Linguistic rules link antecedents to the consequents and are the core of fuzzy inference systems (FIS). At the beginning of the hierarchy, the model used a set of

linguistic rules to capture plausible transformation states and generate the inference necessary within a knowledge base for decision-making. The maximum number of possible linguistic rules based on the exponential function described in Section 3.4.2.3 was six thousand five hundred and sixty-one (3^8 = 6561). To reduce these linguistic rules to manageable numbers, the transformation process was split into two components, namely, the WC-driven and PC-driven transformation processes. As the result at the beginning of the hierarchy, the output of the model was water chemistry-driven aggregation (WCA), water chemistry-driven dissolution (WCD), water chemistry-driven stabilisation (WCS), physicochemical-driven aggregation (PCA), physicochemical-driven dissolution (PCD), and physicochemical-driven stabilisation (PCS).

In each of the developed FIS subsystems at Level-I, the maximum linguistic rules were given by (3 x 3 rule matrix (3^3) = 27). Examples of the developed rules to determine the aggregation from water chemistry and physicochemical parameters are described in Table D.1, respectively. The rule weights of 0.25, 0.50, 0.75, and 1 were assigned to strengthen linguistic rules, and also to resolve conflicting rules when multiple rules apply to a particular case. A total of 108 rules were developed for WC and PC-driven transformation processes. Figure 7.4 shows the surface viewer with the relationship between PC properties and aggregation. Additionally, Figure D.1 shows the fuzzy inference system (FIS) editor that depicts the WCA.

Level II-III. Quantification of transformation processes, e.g. dissolution or aggregation, in an aqueous system is concomitant with high variability and a large standard deviation. These can be linked to a wide range of physicochemical parameters that influence ENP behaviour, variability of instrumentation used, as well as the paucity of standardised experimental procedures (Ban et al., 2018; Basei et al., 2019a). To deal with the high variability of data in quantitative research the normalisation between 0 and 1 has been shown to be quite useful (Yalezo et al., 2024: Yalezo and Musee, 2023). As a result, intermediate parameters e.g. aggregation, and dissolution, among others were defined in the range of 0 to 1 as shown in Table D.2. In this context, zero represents a *very low* degree of aggregation and contrast holds for one.

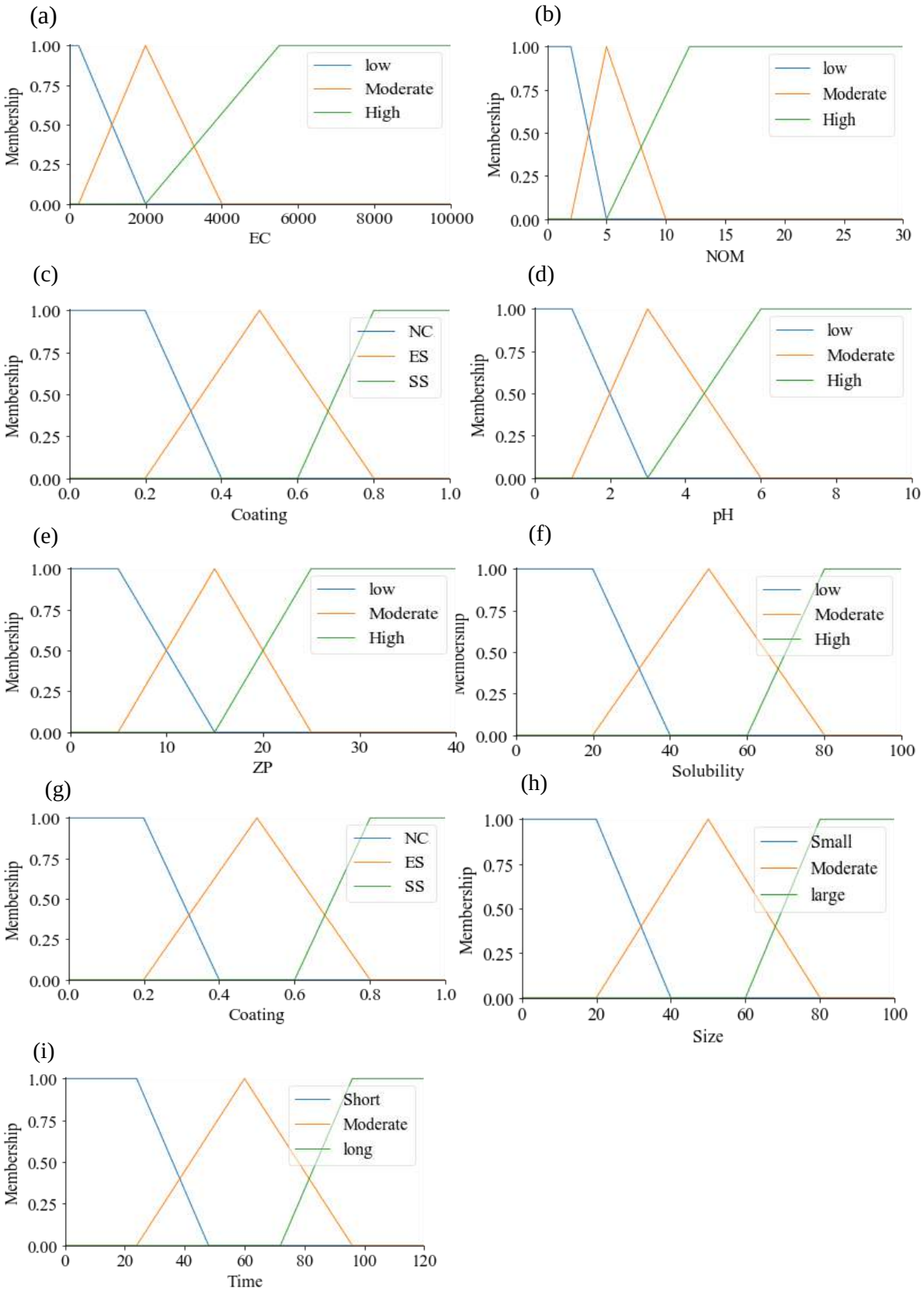


Figure 7. 3. MFs for input parameters

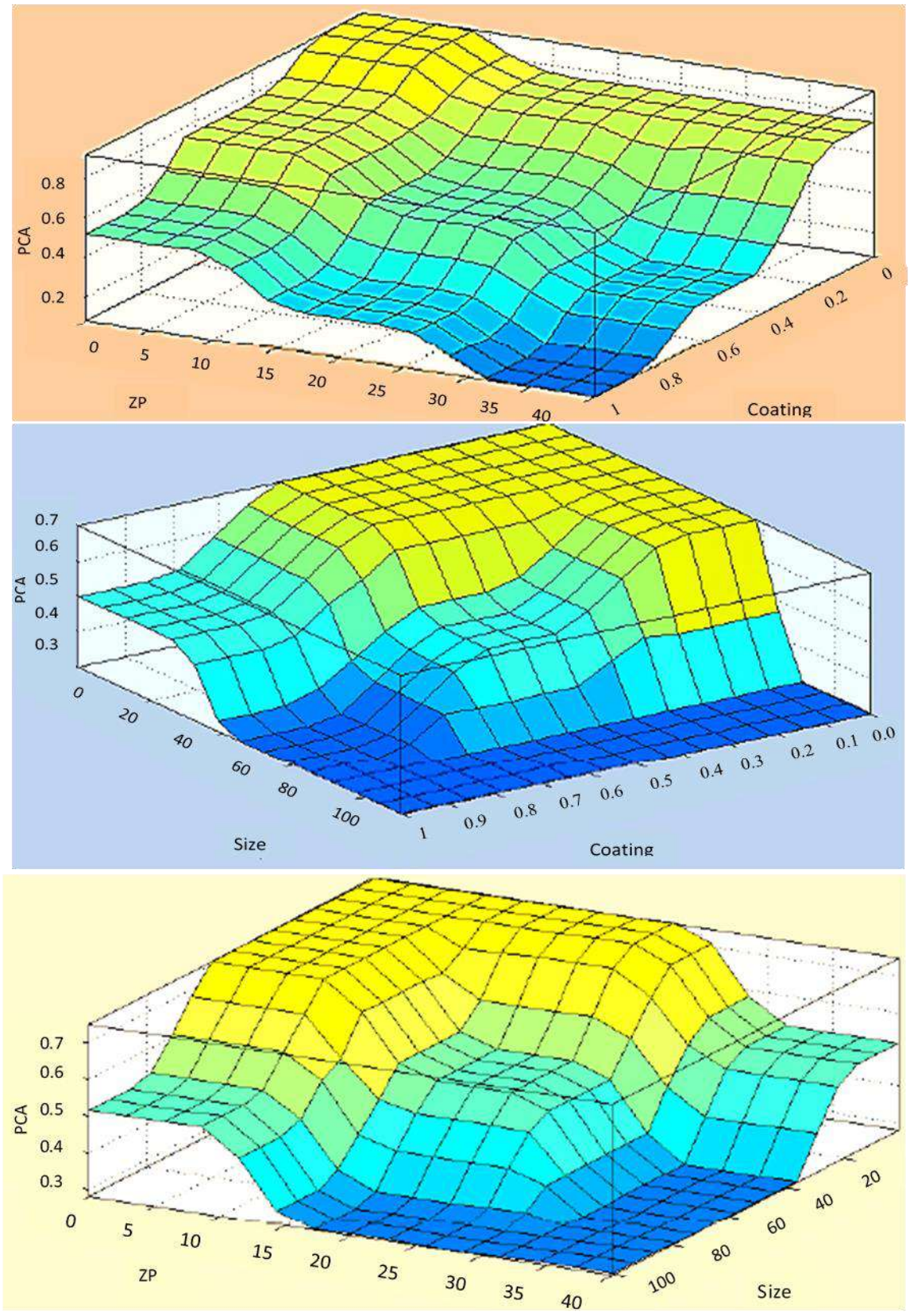


Figure 7. 4. Surface viewer showing the effect of PC properties towards aggregation

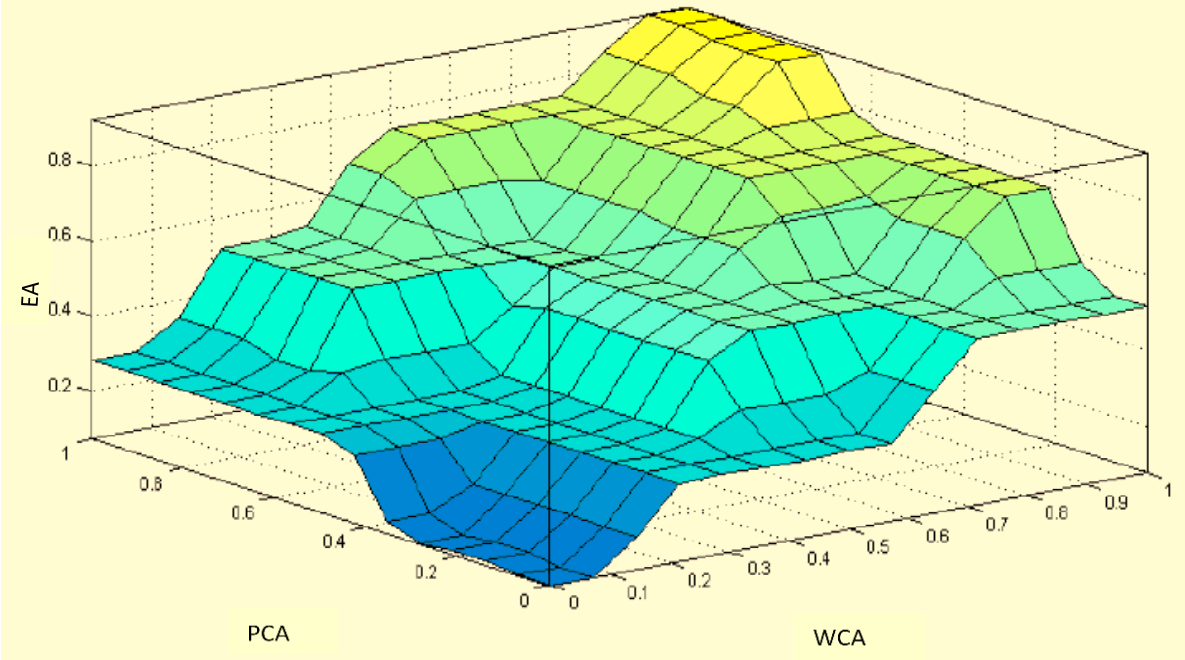


Figure 7. 5. Surface viewer showing the effect of PCA and WCA towards EA.

The system aggregated the WC and PC outputs associated with each transformation process (at level I) to approximate a linguistic label for effective transformations (at level II). These included effective aggregation (EA), effective dissolution (ED), and effective adsorption (EA). The number of rules to estimate effective transformations was described by a 5 x 5 matrix, thus 75 rules could be derived. In the design of if-then rules, WC inputs were assigned higher scores as the preponderant drivers of transformation processes. Figure 7. 5, Figure D.2, and Figure D. 3, show surface viewer, FIS editor, and rule viewer, respectively, depicting a relationship between PCA and WCA with EA. Subsequently, each effective transformation value and the EC index (at Level II) were aggregated to estimate qualitative values for collision efficiency (CE), ionic quantity (IQ) and particulates (PQ) (at Level III). Notable, the particles referred to (at the fourth level of the framework) are modified particles by the WC factors and hence differ from the originally introduced particles. The number of rules to estimate collision efficiency was described by a 5 x 5, whereas particulates and ion quantity were described by a 5 x 5 x 3. This resulted in 175 of the maximum number of rules.

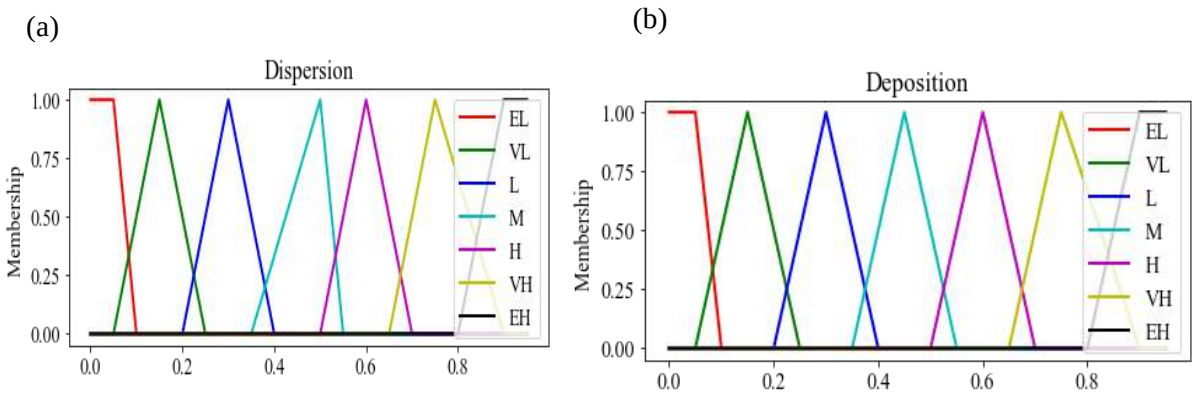


Figure 7. 6. MFs for output parameters

Level-IV: at the end of the hierarchy is the deposition and dispersion MF in 7. 6. The number of rules to determine dispersion and deposition values was described by a matrix of 6 x 6 = 36 and 6 x 3 =18, respectively. To illustrate the fuzzy inference mechanism, suppose the inputs of collision efficiency and exposure are 0.858 and 78 hours, respectively. The inference mechanism controls the manner, in which the rules are executed. The if clause, or precondition, is matched against a series of facts held in the working memory, and the rules that meet the pre-conditions statement are executed and used to produce new sets of facts. The new facts are then matched against other rule preconditions to achieve the solution of the problem to which the rules are designed to apply. For this problem, the four if-then active rules which include rules no. 11, 12, 16 and 17 are aggregated using the *max-min* gravity method as described in Figure 7. 7 as follows.

- Rule 11: If collusion efficiency is *high* AND *Time* is M*edium*
  Then *Deposition* is *high* Evaluation; min (0.28, 0.4) = 0.28
- Rule 12: If collusion efficiency is *very high* AND *Time* is M*edium*
  Then *Deposition* is *Very high*. Evaluation; min (0.53, 0.4) = 0.4
- Rule 16: If collusion efficiency is *high* AND *Time* is High
  Then *Deposition* is *high* Evaluation; min (0.28, 0.27) = 0.27
- Rule 17: If collusion efficiency is *very high* AND *Time* is High
  Then *Deposition* is *very high*. Evaluation; min (0.53, 0.27) = 0.27

Then, using the COG deffuzification method, the crisp output for deposition was estimated as 0.821 and qualitatively classified as *very high* (0.53) and *extremely high* (0.14).

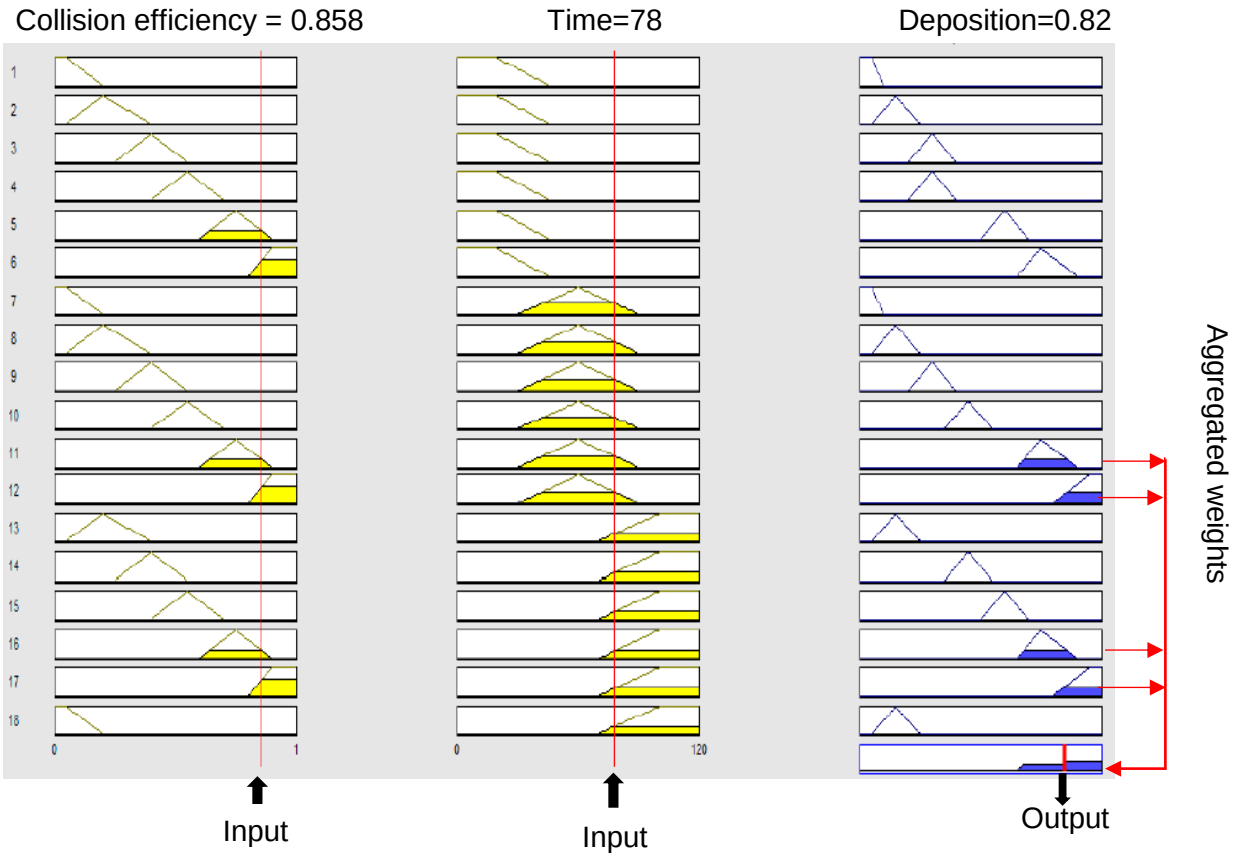


Figure 7. 7. Fuzzy inferencing using Mamdani-Assilian model for the evaluation of ENMs deposition.

7.4 Evaluating the functionality of FDMS

7.4.1 Theoretical Examples

The functionality of the FL model was demonstrated through the use of nZnO and $nTiO_2$ as case studies, because of higher potential exposure to ecological systems (Zhao et al., 2021) as well as their marked difference in solubility (Chen et al., 2016; Dalai et al., 2013).

7.4.2 Results and discussion

Figures 7.8-7.11 and Figures D.4-D.11 systematically illustrate the model functionality of the developed using nZnO and $nTiO_2$. These figures show the user inputs, as well as intermediate and outputs obtained using the developed model. The entries assigned as (–) were treated as zero. The model inputs for nZnO were also made valid for $nTiO_2$. Furthermore, the considered ENPs had fixed values of ZP and size to minimise the complexity of altering many parameters at once. The results obtained from the proposed FL model were argued against the findings previously discussed in multiple reviews and scientific papers. Using both the triangular and trapezoidal MFs the ZP with a crisp input of 21.90 mV and the size 30 nm in the universe of discourse [0, 40] and [0, 100 nm], respectively, were fuzzified as demonstrated below. The ZP was described by three fuzzy sets, namely; μ_{low} (L, 30) = 0, μ_{medium} (M, 13) = 0.31 and μ_{high} (H, 30) = 0.69. Therefore, ZP was linguistically labelled as *moderate - high*, and cannot be linguistically ranked *low*.

$$\mu ZP_{low}(21.90) = \max\left(\min\left(\frac{21.90-0}{0-0}, 1, \frac{15-21.90}{15-5},\right) 0\right)$$

$$= \max(\min\ (1, -0.60), 0)$$

$$= 0$$

$$\mu ZP_{medium}(21.90) = \max\left(\min\left(\frac{21.90-5}{15-5}, \frac{25-21.90}{25-15}\right) 0\right)$$

$$= \max(\min(2.37, 0.315)0)$$

$$= 0.31$$

$$\mu ZP_{high}(21.90) = \max\left(\min\left(\frac{21.90-15}{25-15}, 1, \frac{40-21.90}{40-40}\right), 0\right)$$

$$=\max(\min(0.69), 1)$$

$$= 0.69$$

Furthermore, the size was described by the three fuzzy sets, namely: μ_{small} (L, 30) = 0.33, μ_{medium} (M, 15) = 0.33 and μ_{large} (H, 30) = 0. Therefore, size was labelled *small-medium*, and not in no way can be linguistically ranked as *large*.

$$\mu Size_{low}(30) = \max\left(\min\left(\frac{30-0}{0-0}, 1, \frac{40-30}{40-10},\right) 0\right)$$

$$= \max(\min\ (1, 0.33), 0)$$

= 0.33

$$\mu Size_{medium}(30) = \max\left(\min\left(\frac{30-20}{50-20}, \frac{80-30}{80-50}\right) 0\right)$$

$$= \max(\min(0.33, 1.66,)0)$$

=0.33

$$\mu Size_{high}(30) = \max\left(\min\left(\frac{30-60}{80-60}, 1, \frac{100-30}{100-100}\right), 0\right)$$

$$= \max(\min(-1{,}5, 1), 0)$$

= 0

Table D.3 (nZnO) and Table D.4 ($nTiO_2$) provide the complete set of fuzzy inputs with μ (x) and L_V for all crisp inputs.

In Figure 7.8 (nZnO) the ECs was 32 ng/ℓ and labelled as *low* (1.00). The computed value of PCA was 0.500 and both PCD and PCS were 0.60. PCA was labelled *moderate* (1), and both PCD and PCS were *high* (0.50). The values for WCA, WCD, and WCS were 0.127, 0.327, and 0.550, respectively. The WCA was classified as *very low* (0.73) - *low* (0.14). On the other hand, WCD and WCS were labelled *low* (0.73) and *moderate* (0.50), respectively. Furthermore, the EA, ED and ES were labelled *low* (0.71) - *moderate* (0.04), *low* (0.96) and *moderate* (0.25) - *high* (0.13), respectively, at level II. The deffuzified outputs for deposition and dispersion were 0.146 and 0.450, respectively, at level IV. These were classified as *very low* (0.96) and *moderate* (1.0), correspondingly. Additionally, in Figure 7.9 for $nTiO_2$, a qualitative ranking similar to that discussed for nZnO was derived, despite the marked solubility.

The findings depicted in Figure 7.8 (nZnO) and Figure 7.9 ($nTiO_2$) indicate these ENPs are likely to be dispersed with a minimal degree of deposition. These results were attributed to a NOM of 5.51 mg/ℓ and IS of 2.45 mM that were linguistically labelled *moderate* (0.90) – *high* (0.05) and *low* (0.78) – *moderate* (0.15), respectively. The results observed using a fuzzy system showed consistency with several studies investigating the impact of NOM. According to Abbas et al. (2020) and Louie et al. (2016), the NOM is an electronegative polymer macromolecule that is widely found in the natural system. It constitutes numerous complex biological molecules, including sugars, cellulosic materials, etc. (Aiken et al., 2011; Peng et al., 2015).

A concentration of NOM greater than > 5 mg/ℓ is regarded as high (Leareng et al., 2020). It can adsorb onto ENPs, resulting in the adjustment of physiological PZC, consequently preventing the aggregation of ENPs (Amde et al., 2017; Philippe and Schaumann, 2014). On the other hand, NOM reduces the oxidation of nZnO to toxic ions of Zn^{2+} by blocking oxidation sites; diminishing the bio-accumulation, bio-availability as well as toxicity (Akhil and Sudheer Khan, 2017). For example, dispersion of ENPs due to NOM suppressed the release of ions and the cytotoxic effect of nZnO and nTiO_2 on *alga*l growth (Bhuvaneshwari et al., 2016) and *Chlorella sp* the green algae (Lin et al., 2012), respectively.

Further, in Figure 7.10 and Figure 7.11 identical sets of crisp inputs considered in Figures 7.8 and Figure 7. 9 were applied except the 4.08-fold in IS. The increase in IS resulted in a higher ranking of the deposition for both ENPs in Figure 7.10 and Figure 7.11 compared to Figure 7.8 and Figure 7.9. This was because, according to the literature, increasing IS in the presence of NOM compresses the electric double layer (EDL) with a concomitant increase in the high hetero-aggregation of ENP through cation-induced bridging flocculation or complexation (Leareng et al., 2020). This effect is prominent with divalent electrolytes such as $CaCl_2$ as opposed to NaCl (Chowdhury et al., 2012a). Thus, the findings contained in Figures 7.10-7.11 align with the outcomes documented by Leareng et al. (2020) and Heinlaan et al. (2016). For example, Leareng et al. (2020) and Heinlaan et al. (2016) established that the aggregation behaviour of nZnO and γ-nFe_2O_3 and nCuO was significant in natural water systems that had higher dissolved oxygen-carbon (DOC) and IS, respectively.

In Figures D.4 - D.7 surface coating agents of CIT and PVP resulted in a higher ranking for dispersion. These results complemented previous literature studies as coatings significantly alter the reactivity of ENPs toward many substrates (e.g., NOMs, ions, proteins, etc.) in the aquatic system through the shielding effect (Moore et al., 2015). The sterically charged PVP-coated ENPs in Figure D.4 and Figure D.5 showed a dispersion ranking higher than that of the CIT-coated ones in Figure D.6 and Figure D.7. Differences in the degree of stabilisation were explained by the high molecular weight of PVP-coated ENPs (Ellis et al., 2016).

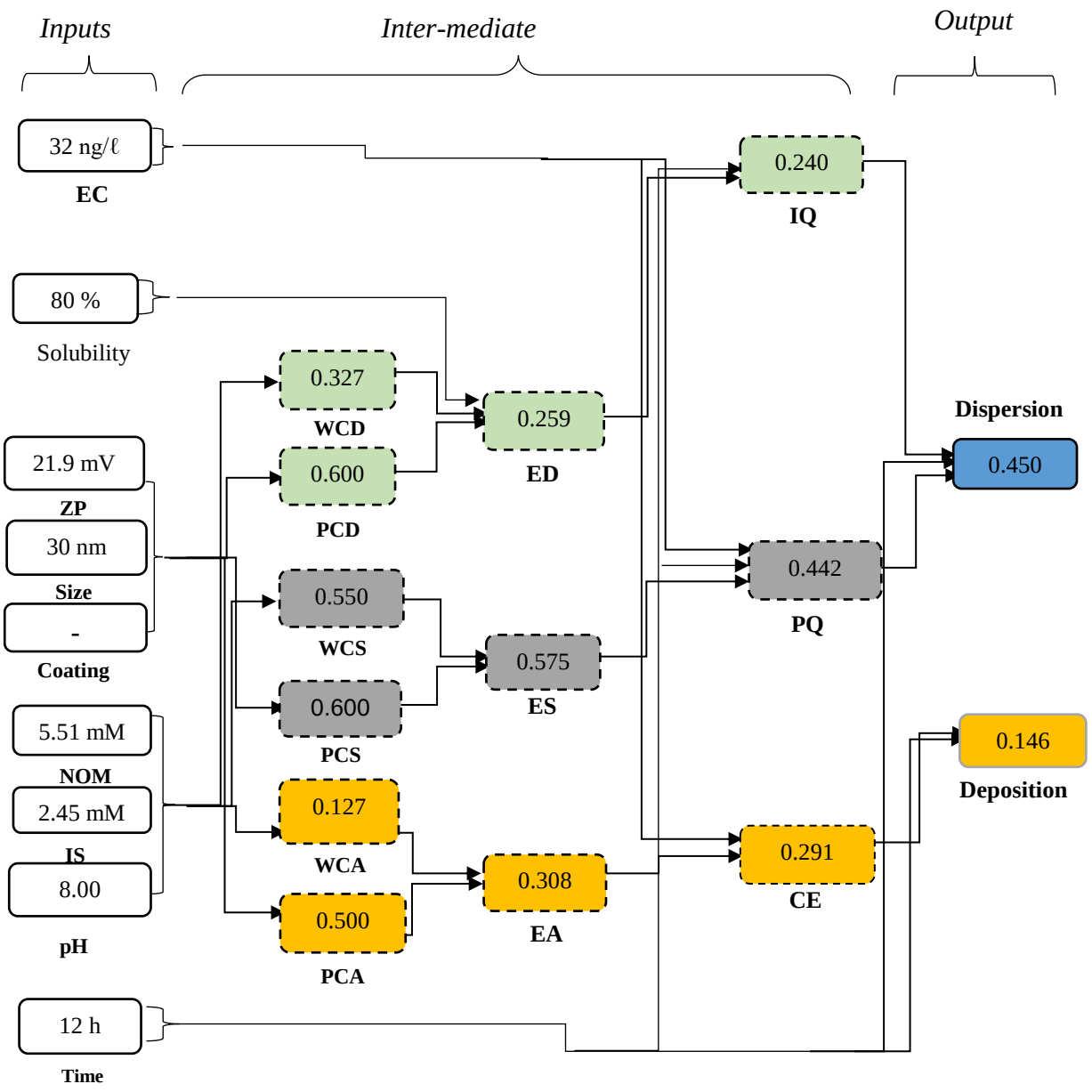


Figure 7. 8. Illustrating stepwise functionality of FL using nZnO for Scenario 1

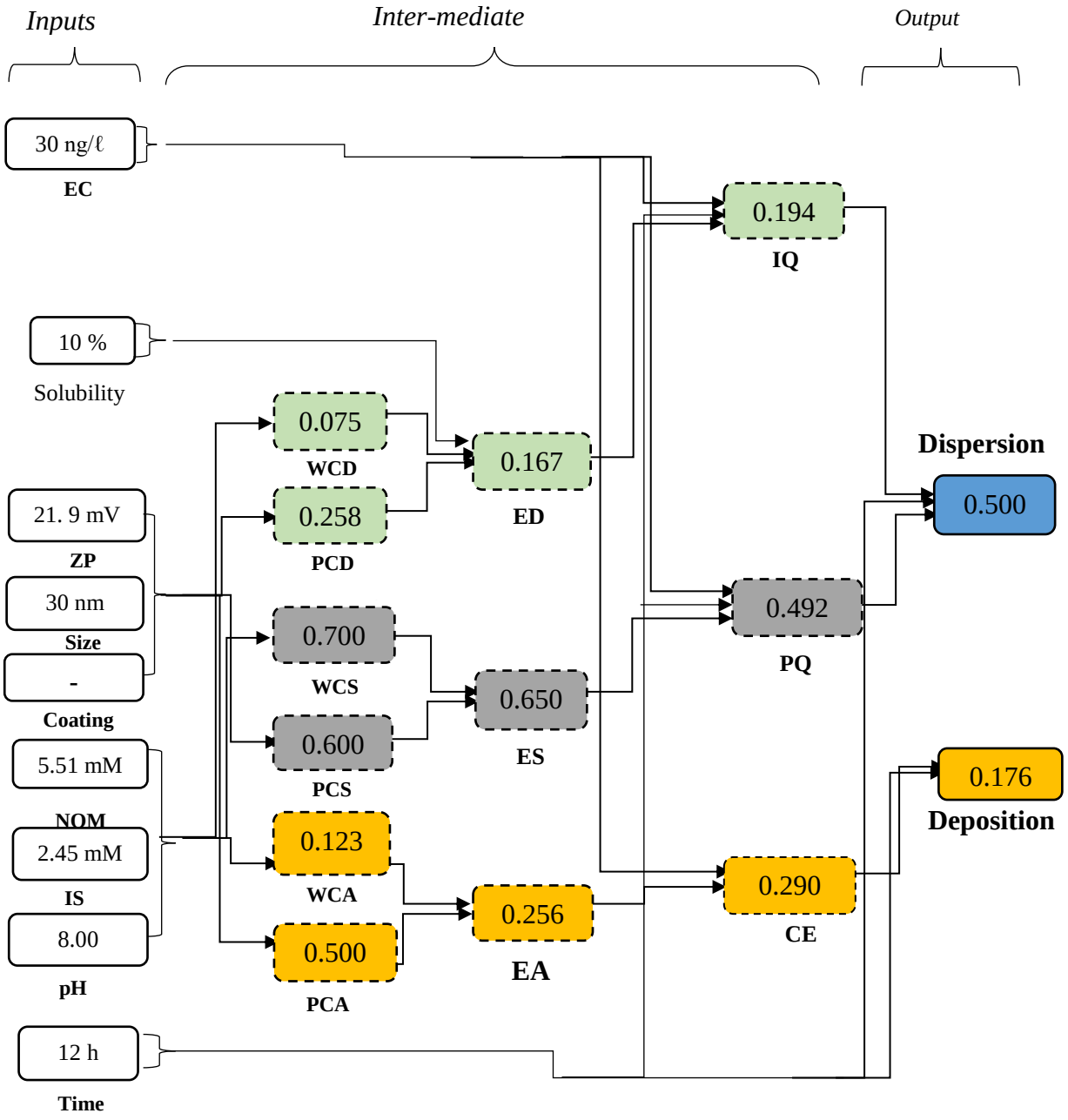


Figure 7. 9. Illustrating stepwise functionality of FL using $nTiO_2$ for Scenario 1

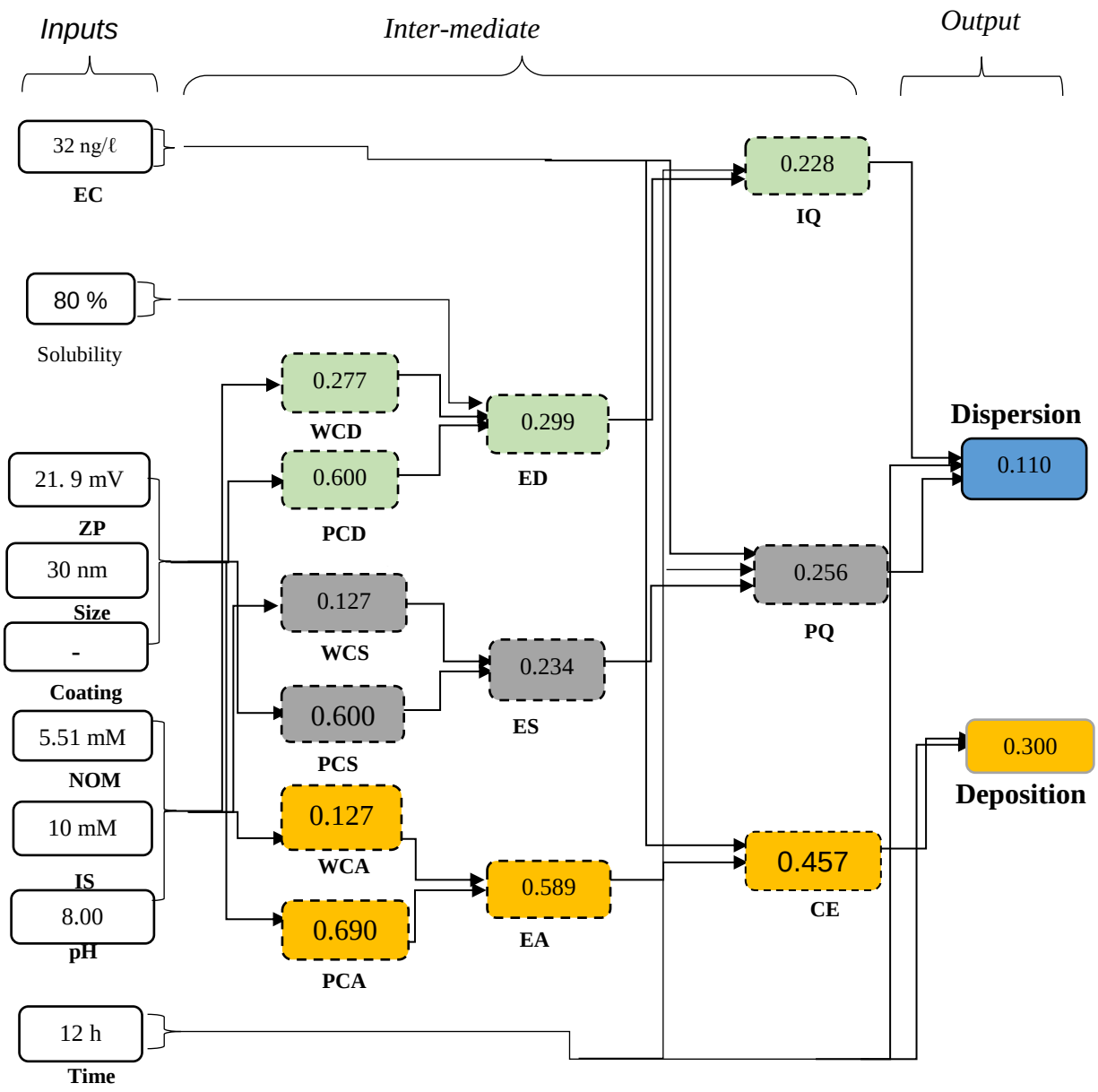


Figure 7. 10. Illustrating stepwise functionality of FL using nZnO for Scenario 2

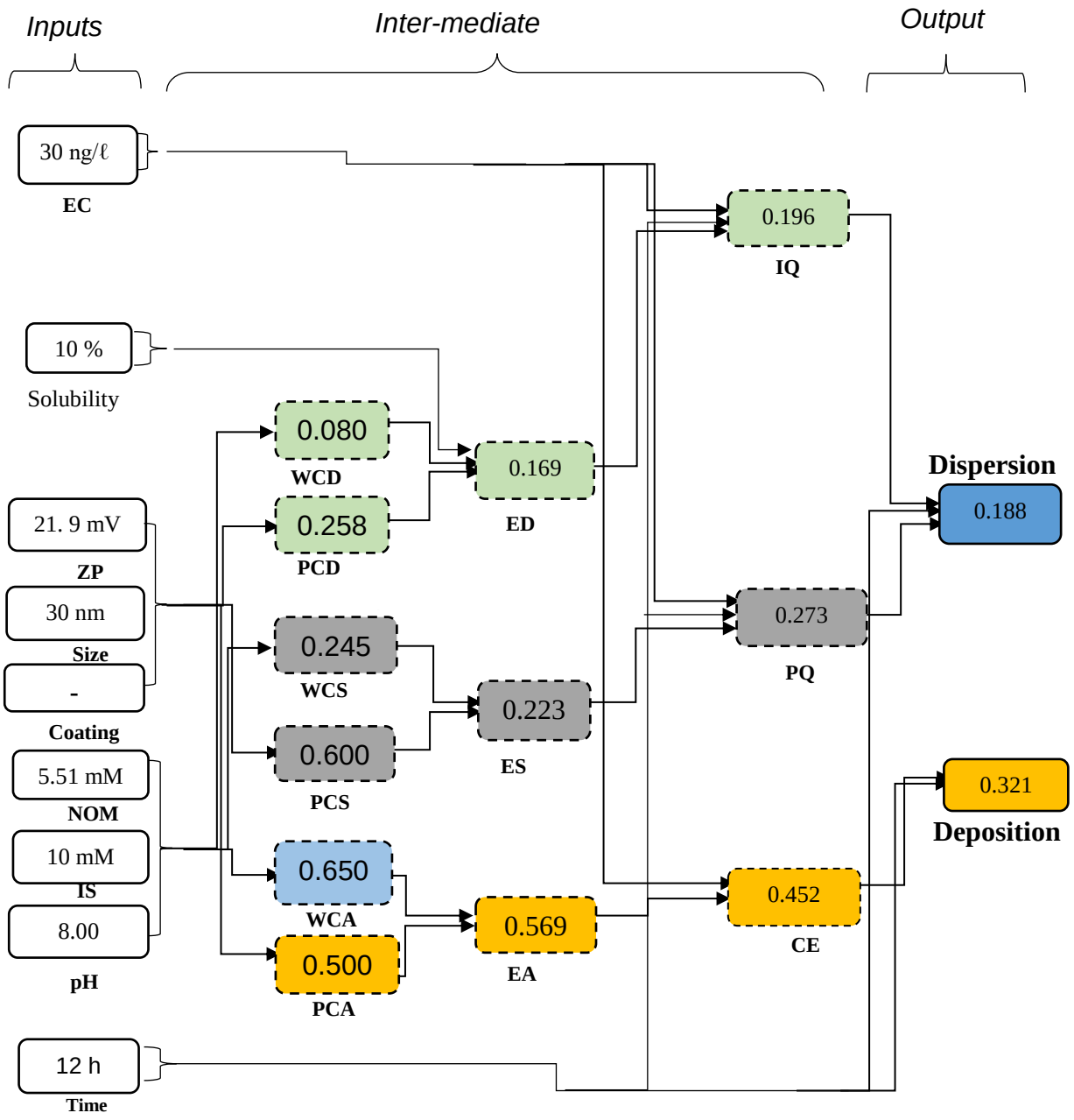


Figure 7. 11. Illustrating stepwise functionality of FL using $nTiO_2$ for Scenario 2

Furthermore, the model predicted that the nZnO in Figure D.8 and Figure D.10 undergoes a higher degree of dissolution and reduced aggregation, while $nTiO_2$ in Figure D.9 and Figure D.11 undergoes a higher degree of aggregation, which can be associated with the PZC regime. The PZC is a regime of weak electrostatic repulsion forces and strong van der Waals (vdW), resulting in maximum aggregation (Abbas et

al., 2020; Loosli et al., 2015). The PZC is found in the pH region of 8.5 to 9.0 for nZnO (Peng et al., 2015) and 5-6 for $nTiO_2$ (Loosli et al., 2015). Furthermore, the 16-fold increase in exposure time resulted in a higher ranking of the deposition of both ENPs between Figure D.9 and Figure D.11. According to the Stokes theorem, longer exposure times are related to larger clusters of aggregates (Schaumann et al., 2015). For example, Dalai et al. (2013) demonstrated that a longer exposure time corresponded to improved agglomeration-sedimentation rates and reduction in $nTiO_2$ toxicity in the freshwater algal isolate *Scenedesmus obliquus*.

7.5 Model generalisation and limitation

The fuzzy model developed herein can be highly valuable as the first screening tool for the transformation and exposure of ENPs, and this can aid in chemical characterisation and monitoring of the exposure of ENPs in aquatic systems. The fuzzy system followed a hierarchical framework constituting eight deterministic antecedent parameters (size, zeta potential, NOM, IS, time, pH, concentration of ENPs per given nano-products, use frequency) and three qualitative (surface coating, solubility, product matrix), linked with intermediate parameters and exposure outputs indicators of dispersion and deposition. The developed fuzzy model took into account the degree of influence of various inputs. About 321 (three hundred and twenty-one) linguistic rules were coded to connect the antecedents to the consequents.

From the results the developed fuzzy logic system was able to generalise both ENPs and the approach presented here illustrates an integrated framework that can aid in simultaneously assessing the transformation processes, as the model was shown as effective across ENPs with different solubilities. The developed fuzzy decision system is less dependent on big numerical data compared to machine learning, as well as user-friendly and easily interpretable by non-technical experts (Di Addario et al., 2017; Paul et al., 2018). Furthermore, different metal-based ENPs need to be investigated to show the flexibility of the proposed model. In addition, work examples developed to demonstrate the functionality are not exhaustive, more different metal-based ENPs and other aquatic permutations to demonstrate the flexibility of the proposed model are necessary. The efficacy and accuracy of the proposed fuzzy logic system to assess ENP exposure is hinged on a knowledge library that constitutes linguistic rules that link the antecedents to the consequent.

Additionally, the Mamdani-Assilian fuzzy model was successful in integrating uncertainty and was best suited for interpretation; however, challenges include (1) knowledge gathering or expert information process, (2) model development, and (3) generalisation of the application domain. The first challenge rises from various interpretations of knowledge that can induce a personal bias in perception. In certain circumstances, expert opinions my lack the level of consistency and different experts may propose dissimilar or contradictory decisions.

The second challenge is the assignment of fuzzy sets and the process of developing consequences of linguistic rules, which can be tedious and prone to human error (Paul et al., 2018; Zhang et al., 2013; Zhou et al., 2022). Linguistic rules have a major influence on the model findings as they encapsulate the heuristic knowledge of the domain to generate the knowledge-rule base necessary for decision-making (Paul et al., 2018). As described in Section 3.4.2.3, knowledge is represented using the linguistic rules composed of syntax in the format of "*if* "(situation, condition, pattern) and "*then"* actions are utilised to link identified inputs to outputs (Amiri et al., 2017; Araya-Muñoz et al., 2017). Theoretically, the maximum number of permissible fuzzy rules in the rule editor interface is described by an exponential function depending on the designated input and MFs. As a result, the number of fuzzy rules in the fuzzy inference systems increases exponentially relating to the number of input variables used (Chiu, 1996; Shi et al., 1999; Zolghadri et al., 2007). Therefore, high numbers of rules can result in an unmanageable number of linguistic rules and prone to human error.

Third, fuzzy models are static to expert knowledge used and not highly generalised to detect out-of-sample information. The level of accuracy and generalisation may only be within the confines of data insights gathered or encoded as the model can become ineffective to generalise values outside the upper or lower limits. Therefore, as the new information becomes accessible and considered relevant in the freshwater system the developed 340 if-then linguistic rules systematically encoded following the hierarchical framework described in Figure 7.2 will require further manual updates. Thus, the authors acknowledge that this task is subjective and challenging due to expert knowledge, yet it forms the core step of the FIS, and in turn, has an influence on the model findings. Therefore, as quantitative data in this domain become more

readily available and standardised, in future work, there is a need for the use of dynamic techniques such as adaptive neuro-fuzzy systems and K-means clustering to automate fuzzification and development of linguistic rules (Jansson et al., 2022; Mirzakhanov, 2020).

7.6 Environmental Significance and Model Deployment

The continuous emissions of engineered nanoparticles (ENPs) into aquatic systems have raised serious concerns about potential deleterious effects on biological lifeforms. Evaluation of ENP exposure and risk assessment in aquatic systems requires quantification of their emission, but more importantly, understanding of their colloidal stability, which drives accumulation and bioavailability. Presently, information regarding the bioavailability and accumulation of ENPs at various tropic levels of the aquatic systems is characterized by knowledge gaps, and data uncertainty, which, renders it challenging to support decision-making. On the other hand, the existing approaches are characterised by large data uncertainties which consequently, limits their informative aspects essential for the decision-making environment (Hristozov et al., 2016; Nowack et al., 2015). For example, the semi-quantitative models use Boolean systems, therefore, do not address the data uncertainty, imprecision, and ambiguity, that arises with the modelling of expert intuitions or perceptions (Ye et al., 2020).

Predicting the potential fate and transformation of ENPs in the natural system can be essential to determine their environmental effect (Abbas et al., 2020). Thus, the developed fuzzy model herein can be highly valuable as the first screening tool for the transformation and exposure of ENPs, and this can aid in chemical characterization and monitoring of the exposure of ENPs in aquatic systems. The fuzzy logic approach provided rational and well-worked solutions and was suitable to handle human cognitive processes, common sense knowledge, and linguistic data, therefore, can make it feasible to deduce definite conclusions (Topuz et al., 2016; Iancu, 2019). The practical application of the developed model is through a user interface. However, at present, these systems are being developed and will be validated by experts before deployment for accessibility to users.

7.7 Chapter summary

This chapter has demonstrated the suitability of the proposed FDMS as a decision tool to screen the ENP exposure and likely environmental impact in aquatic systems. Fuzzy logic showed suitability as a computational tool in the domain defined with uncertainty and ambiguity for estimating the exposure potential of ENPs. This is because in this case the cause and effect due to inherent ENP physicochemical properties and the abiotic factors with respect to; fate, behaviour, and toxicity in different ecosystems are poorly understood. The framework combined both quantitative and qualitative information into a single framework. To the author's knowledge, neither previously experimental nor modelling studies have reported both the predicted ENP concentration emissions, with fate and behavior in single research papers. Therefore, this approach presents a new integrated framework that can aid in simultaneously assessing the three transformation processes. In addition, the developed model can be easily optimised as new information becomes available, without the need to reconstruct the entire model. These findings will motivate further application of FL for addressing numerous aspects within nanoecotoxicology for a wide range of ENPs.

Chapter 8. Conclusions, drawbacks, and recommendations

8.1 Conclusions

In conclusion, this work has successfully developed various modelling by integration of data-driven and knowledge-based approaches to provide a foundation for the initial screening and monitoring of ENP exposure and risk assessment. This research successfully combined data-driven ML techniques with rule-based KBS to maximize the utility of both structured experimental data and unstructured expert knowledge. This hybrid approach provided a scalable, adaptable framework for continuous monitoring, initial screening, and risk assessment of ENPs, thereby reducing the reliance on costly and time-consuming experimental testing. The developed models not only facilitate nano-safety evaluations but also offer a foundation for extending these techniques to other classes of ENPs, promoting sustainable advancements in nanotechnology while ensuring long-term environmental protection. The developed model advances computational techniques, contributing to both academic knowledge and practical applications in environmental engineering and nanotechnology.

(i) Development of robust machine learning models for predicting ENP behavior

ML algorithms of RF, XGBoost, SVR and ANN produced a satisfactory level of accuracy in predicting the aggregation or dissolution of nZnO and $nTiO_2$. On the contrary, MLR models showed relatively large RMSE and MAE for both ENPs and different transformations of aggregation and dissolution. In addition, to ML's exceptional potential as a predicting tool, it has guided experimental investigations by providing a deeper understanding of significant physicochemical descriptors, thus; reducing the focus on a large number of predictors to a small number of selected variables.

Input parameters of NOM, IS, size, and ENP concentration were found to have a poor prediction of the dynamic aggregation of nZnO, and $nTiO_2$ in aqueous systems despite experimentally being identified as significant. The input parameter of ZP of ENPs, the pH of the aqueous system, and the duration of exposure (time) were identified as predominant and are more reliable parameters that can help simulate and predict ENPs' aggregation of ENPs in freshwater-like media based on the heterogeneous data

sets analysed. On the other hand, the continuous inputs including the NOM, time, ENP concentration, size, IS, and pH showed good generalisation and accuracy in estimating the concentration of Zn^{2+}. In addition, categorical input variables including coating, coating type, and NOM type had low predictive power.

(ii) Development semi-quantitative analysis and fuzzy decision-making systems for intelligent decision-making

To account for both qualitative and quantitative data the computer-based KBS were successfully developed based on the parsimonious theoretical conceptual framework. These are intelligent KBS such as the semi-quantitative analysis with decision tree classifiers and fuzzy decision-making systems. The multi-variant factors influencing the exposure output of dispersion stability and deposition were mapped using information structures of MISO (Multiple Input, Single Output). These systems effectively encoded expert domain knowledge into structured decision-making frameworks, enabling cost-effective, user-friendly, and non-expert-accessible tools for evaluating the environmental impact of ENPs. Results demonstrated that a computer-based KBS can integrate ill-defined variables, non-linear data, and a paucity of quantitative structured data, expert knowledge, intuition, and heuristics in the process of decision-making. They balance the simplicity of qualitative approaches with the increased detail of quantitative methods. They allow for efficient, flexible, and practical assessments, especially in scenarios where full quantitative data or analyses are not feasible. Thus, can be highly valuable as a screening tool in the domain defined by a lack of sharp boundaries.

8.2 Drawbacks and Challenges

The study objectives have been successfully addressed; however, several drawbacks were identified in the data or application of ML. First, the accessible data are defined by inconsistent reporting protocols including influencing factors, exposure media, high variability of instruments used for measuring various parameters, and deficiency of appropriate controls, despite from early years of ENP nano-safety investigations where the urgent need for minimum data reporting protocols was raised (Holden et al., 2016; Pettitt and Lead, 2013). The use of both the synthetic media and freshwater system presents huge uncertainty and bias in the developed model as it might

compromise or weaken the strength of variables. Other synthetic-based studies use little or no presence of NOM concentration.

NOM and IS are reported as static variables in experimental settings (Monikh et al., 2018) and, may not be reflective of the actual dynamic environmental systems of aggregation and dis-agglomeration processes that occur continuously during prolonged exposure. The qualitative data such as the type of NOM presents obvious complications in the context of modelling the fate and transport of ENPs. Variant types of NOM have different molecular weight distributions, which in turn are likely to have a different extent of influence on ENP aggregation (Louie et al., 2016, 2013). The majority of research studies used dynamic light scattering (DLS) to measure HDD, but DLS has significant limitations (i.e., it cannot measure polydisperse samples) which may result in bias (Mahl et al., 2011; Uskoković, 2012).

Moreover, the application of meta-analysis allows quantitative evaluation of highly heterogeneous data from multiple sources (Gurevitch, 1993), but the approach has several pitfalls including the management of publication bias in favour of positive results, the lack of randomised sampling, and strong dependence on the quality of studies included for analysis (Greco et al., 2013; Walker et al., 2008). It is not feasible to find every relevant study on a subject. Many studies may not be published, or indexed in computer-searchable databases.

8.3 Recommendations

In future study, the developed model's accuracy and model resolution can be refined through the investigation of specific types of ENPs within a specific class as well. For example, herein all $nTiO_2$ were considered as one group of ENPs; yet $nTiO_2$-rutile and $nTiO_2$-anatase are known to exhibit different aggregation profiles (Liu et al., 2011). The use of natural systems should be the most viable approach since it would provide a better picture of the actual interaction between natural colloidal particles and ENPs. In addition, future studies should also focus on the investigation of aged material as opposed to pristine. The likelihood of pristine ENPs being released into the environment is low. Instead, aged ENPs that have substantially different physiochemical properties than their pristine counterparts are more likely to be released into the environment (Mitrano and Nowack, 2017).

Suwannee River NOM (SR-NOM) should be used uniformly as a surrogate for NOM to improve the reproducibility of the results (Monikh et al., 2018). This is because the success of any modelling not only depends on the technique(s) employed but also on the quality and accuracy of the data used. Therefore, a standardised experimental process and improvements in methodology are required. This can be critical for the derivation of the cut-off values and reproducibility. Furthermore, future studies based on meta-analysis should consider techniques and approaches that address meta-analysis limitations including the calculation of studies needed to refute or affirm conclusions based on a meta-analysis (Rosenthal, 1979). Re-addressing these limitations can aid to improve the quality of data and provide better insights valuable to decision- and policy-formulation. In addition for practical application, the is a need to develop web based website and deploy it in a cloud system to allow for universal usage.

Reference

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. Environment international 138, 105646.

Abdolahpur Monikh, F., Praetorius, A., Schmid, A., Kozin, P., Meisterjahn, B., Makarova, E., Hofmann, T., von der Kammer, F., 2018. Scientific rationale for the development of an OECD test guideline on engineered nanomaterial stability. NanoImpact 11, 42–50. https://doi.org/10.1016/j.impact.2018.01.003

Adam, V., Loyaux-Lawniczak, S., Labille, J., Galindo, C., del Nero, M., Gangloff, S., Weber, T., Quaranta, G., 2016. Aggregation behaviour of TiO2 nanoparticles in natural river water. J Nanopart Res 18, 13. https://doi.org/10.1007/s11051-015-3319-4

Adam, V., Nowack, B., 2017. European country-specific probabilistic assessment of nanomaterial flows towards landfilling, incineration and recycling. Environ. Sci.: Nano 4, 1961–1973. https://doi.org/10.1039/C7EN00487G

Agatonovic-Kustrin, S., Beresford, R., 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of Pharmaceutical and Biomedical Analysis 22, 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

Ahmad, F., Mat Isa, N.A., Hussain, Z., Osman, M.K., 2013. Intelligent medical disease diagnosis using improved hybrid genetic algorithm-multilayer perceptron network. Journal of medical systems 37, 1–8.

Aiken, G.R., Hsu-Kim, H., Ryan, J.N., 2011. Influence of dissolved organic matter on the environmental fate of metals, nanoparticles, and colloids. ACS Publications.

Akhil, K., Sudheer Khan, S., 2017. Effect of humic acid on the toxicity of bare and capped ZnO nanoparticles on bacteria, algal and crustacean systems. Journal of Photochemistry and Photobiology B: Biology 167, 136–149. https://doi.org/10.1016/j.jphotobiol.2016.12.010

Akobeng, A.K., 2005. Principles of evidence based medicine. Archives of disease in childhood 90, 837–840.

Akter, M., Sikder, M.T., Rahman, M.M., Ullah, A.A., Hossain, K.F.B., Banik, S., Hosokawa, T., Saito, T., Kurasaki, M., 2018. A systematic review on silver

nanoparticles-induced cytotoxicity: Physicochemical properties and perspectives. Journal of advanced research 9, 1–16.

Alade, I.O., Abd Rahman, M.A., Abbas, Z., Yaakob, Y., Saleh, T.A., 2020. Application of support vector regression and artificial neural network for prediction of specific heat capacity of aqueous nanofluids of copper oxide. Solar Energy 197, 485–490.

Alade, I.O., Abd Rahman, M.A., Saleh, T.A., 2019. Modeling and prediction of the specific heat capacity of Al2 O3/water nanofluids using hybrid genetic algorithm/support vector regression model. Nano-Structures & Nano-Objects 17, 103–111.

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

Algar, W.R., Tavares, A.J., Krull, U.J., 2010. Beyond labels: a review of the application of quantum dots as integrated components of assays, bioprobes, and biosensors utilizing optical transduction. Analytica chimica acta 673, 1–25.

Alharbi, O.M., Khattab, R.A., Ali, I., 2018. Health and environmental effects of persistent organic pollutants. Journal of Molecular Liquids 263, 442–453.

Ali, O.A.M., Ali, A.Y., Sumait, B.S., 2015. Comparison between the Effects of Different Types of Membership Functions on Fuzzy Logic Controller Performance.

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

Alin, A., 2010. Multicollinearity: Multicollinearity. WIREs Comp Stat 2, 370–374. https://doi.org/10.1002/wics.84

Al-Kattan, A., Wichser, A., Vonbank, R., Brunner, S., Ulrich, A., Zuin, S., Arroyo, Y., Golanski, L., Nowack, B., 2015. Characterization of materials released into water from paint containing nano-SiO2. Chemosphere 119, 1314–1321. https://doi.org/10.1016/j.chemosphere.2014.02.005

Alkilany, A.M., Mahmoud, N.N., Hashemi, F., Hajipour, M.J., Farvadi, F., Mahmoudi, M., 2016. Misinterpretation in nanotoxicology: A personal perspective. Chemical research in toxicology 29, 943–948.

Alpaydin, E., 2020. Introduction to machine learning. MIT press.

Ambure, P., Ballesteros, A., Huertas, F., Camilleri, P., Barigye, S.J., Gozalbes, R., 2020. Development of Generalized QSAR Models for Predicting Cytotoxicity and Genotoxicity of Metal Oxides Nanoparticles: International Journal of Quantitative Structure-Property Relationships 5, 83–100. https://doi.org/10.4018/IJQSPR.20201001.oa2

Amde, M., Liu, J., Tan, Z.-Q., Bekana, D., 2017. Transformation and bioavailability of metal oxide nanoparticles in aquatic and terrestrial environments. A review. Environmental Pollution 230, 250–267. https://doi.org/10.1016/j.envpol.2017.06.064

Amiri, M., Ardeshir, A., Fazel Zarandi, M.H., 2017. Fuzzy probabilistic expert system for occupational hazard assessment in construction. Safety Science 93, 16–28. https://doi.org/10.1016/j.ssci.2016.11.008

Amirshenava, S., Osanloo, M., 2019. A hybrid semi-quantitative approach for impact assessment of mining activities on sustainable development indexes. Journal of Cleaner Production 218, 823–834. https://doi.org/10.1016/j.jclepro.2019.02.026

Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural computation 9, 1545–1588.

An, H., Zhou, Z., Yi, Y., 2017. Opportunities and challenges on nanoscale 3D neuromorphic computing system, in: 2017 IEEE International Symposium on Electromagnetic Compatibility & Signal/Power Integrity (EMCSI). IEEE, pp. 416–421.

Aqlan, F., 2016. A software application for rapid risk assessment in integrated supply chains. Expert Systems with Applications 43, 109–116. https://doi.org/10.1016/j.eswa.2015.08.028

Aquilina, N.J., Delgado-Saborit, J.M., Bugelli, S., Ginies, J.P., Harrison, R.M., 2018. Comparison of Machine Learning Approaches with a General Linear Model To Predict Personal Exposure to Benzene. Environ. Sci. Technol. 52, 11215–11222. https://doi.org/10.1021/acs.est.8b03328

Araya-Muñoz, D., Metzger, M.J., Stuart, N., Wilson, A.M.W., Carvajal, D., 2017. A spatial fuzzy logic approach to urban multi-hazard impact assessment in Concepción, Chile. Science of The Total Environment 576, 508–519. https://doi.org/10.1016/j.scitotenv.2016.10.077

Arts, J.H., Hadi, M., Irfan, M.-A., Keene, A.M., Kreiling, R., Lyon, D., Maier, M., Michel, K., Petry, T., Sauer, U.G., 2015. A decision-making framework for the grouping and testing of nanomaterials (DF4nanoGrouping). Regulatory Toxicology and Pharmacology 71, S1–S27.

Arvidsson, R., Molander, S., Sandén, B.A., Hassellöv, M., 2011. Challenges in Exposure Modeling of Nanoparticles in Aquatic Environments. Human and Ecological Risk Assessment: An International Journal 17, 245–262. https://doi.org/10.1080/10807039.2011.538639

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

Balsara, S., Jain, P.K., Ramesh, A., 2019. An integrated approach using AHP and DEMATEL for evaluating climate change mitigation strategies of the Indian cement manufacturing industry. Environmental pollution 252, 863–878.

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. U.S.A. 117, 10492–10499. https://doi.org/10.1073/pnas.1919755117

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

Basei, G., Hristozov, D., Lamon, L., Zabeo, A., Jeliazkova, N., Tsiliki, G., Marcomini, A., Torsello, A., 2019a. 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

Basei, G., Hristozov, D., Lamon, L., Zabeo, A., Jeliazkova, N., Tsiliki, G., Marcomini, A., Torsello, A., 2019b. 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. Applied artificial intelligence 17, 519–533.

Batista, G.E., Monard, M.C., 2002. A Study of K-Nearest Neighbour as an Imputation Method. His 87, 48.

Baun, A., Sayre, P., Steinhäuser, K.G., Rose, J., 2017. Regulatory relevant and reliable methods and data for determining the environmental fate of manufactured nanomaterials. NanoImpact 8, 1–10. https://doi.org/10.1016/j.impact.2017.06.004

Belvederesi, C., Dominic, J.A., Hassan, Q.K., Gupta, A., Achari, G., 2020. Predicting river flow using an AI-based sequential adaptive neuro-fuzzy inference system. Water 12, 1622.

Bennajeh, A., Bechikh, S., Said, L.B., Aknine, S., 2018. A Fuzzy Logic-Based Anticipation Car-Following Model, in: Thanh Nguyen, N., Kowalczyk, R. (Eds.), Transactions on Computational Collective Intelligence XXX, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 200–222. https://doi.org/10.1007/978-3-319-99810-7_10

Besinis, A., De Peralta, T., Handy, R.D., 2014. The antibacterial effects of silver, titanium dioxide and silica dioxide nanoparticles compared to the dental disinfectant chlorhexidine on Streptococcus mutans using a suite of bioassays. Nanotoxicology 8, 1–16.

Bhuvaneshwari, M., Iswarya, V., Nagarajan, R., Chandrasekaran, N., Mukherjee, A., 2016. Acute toxicity and accumulation of ZnO NPs in Ceriodaphnia dubia: Relative contributions of dissolved ions and particles. Aquatic Toxicology 177, 494–502. https://doi.org/10.1016/j.aquatox.2016.07.003

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

Biau, G., 2012. Analysis of a random forests model. The Journal of Machine Learning Research 13, 1063–1095.

Biau, G., Devroye, L., 2010. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. Journal of Multivariate Analysis 101, 2499–2518. https://doi.org/10.1016/j.jmva.2010.06.019

Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial intelligence 97, 245–271.

Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K., 1987. Occam's razor. Information processing letters 24, 377–380.

Bockstaller, C., Beauchet, S., Manneville, V., Amiaud, B., Botreau, R., 2017. A tool to design fuzzy decision trees for sustainability assessment. Environmental Modelling & Software 97, 130–144. https://doi.org/10.1016/j.envsoft.2017.07.011

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., 2013. A review of feature selection methods on synthetic data. Knowl Inf Syst 34, 483–519. https://doi.org/10.1007/s10115-012-0487-8

Bossa, N., Chaurand, P., Levard, C., Borschneck, D., Miche, H., Vicente, J., Geantet, C., Aguerre-Chariol, O., Michel, F.M., Rose, J., 2017. Environmental exposure to TiO2 nanomaterials incorporated in building material. Environmental Pollution 220, 1160–1170. https://doi.org/10.1016/j.envpol.2016.11.019

Bouchrika, T., Jemai, O., Zaied, M., Ben Amar, C., 2014. A New Hand Posture Recognizer Based on Hybrid Wavelet Network Including a Fuzzy Decision Support System, in: Corchado, E., Lozano, J.A., Quintián, H., Yin, H. (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2014, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 183–190. https://doi.org/10.1007/978-3-319-10840-7_23

Braae, M., Rutherford, D.A., 1978. Fuzzy relations in a control setting. Kybernetes.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Breiman, L., 1996. Bagging predictors. Mach Learn 24, 123–140. https://doi.org/10.1007/BF00058655

Brubaker, R., 1985. Rethinking classical theory. Theory and society 14, 745–775.

Brunelli, A., Pojana, G., Callegaro, S., Marcomini, A., 2013. Agglomeration and sedimentation of titanium dioxide nanoparticles (n-TiO2) in synthetic and real waters. J Nanopart Res 15, 1684. https://doi.org/10.1007/s11051-013-1684-4

Bundschuh, M., Filser, J., Lüderwald, S., McKee, M.S., Metreveli, G., Schaumann, G.E., Schulz, R., Wagner, S., 2018. Nanoparticles in the environment: where do we come from, where do we go to? Environ Sci Eur 30, 6. https://doi.org/10.1186/s12302-018-0132-6

Buragohain, M., Mahanta, C., 2008. A novel approach for ANFIS modelling based on full factorial design. Applied soft computing 8, 609–625.

Burney, S.M.A., Jilani, T.A., Ardil, C., 2007. A comparison of first and second order training algorithms for artificial neural networks. International Journal of Computer and Information Engineering 1, 145–151.

Buzea, C., Pacheco, I.I., Robbie, K., 2007. Nanomaterials and nanoparticles: sources and toxicity. Biointerphases 2, MR17–MR71.

Camastra, F., Ciaramella, A., Giovannelli, V., Lener, M., Rastelli, V., Staiano, A., Staiano, G., Starace, A., 2015. A fuzzy decision system for genetically modified plant environmental risk assessment using Mamdani inference. Expert Systems with Applications 42, 1710–1716. https://doi.org/10.1016/j.eswa.2014.09.041

Cameron, S.J., Hosseinian, F., Willmore, W.G., 2018. A current overview of the biological and cellular effects of nanosilver. International journal of molecular sciences 19, 2030.

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.

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

Castillo, O., Melin, P., 2014. A review on interval type-2 fuzzy logic applications in intelligent control. Information Sciences 279, 615–631. https://doi.org/10.1016/j.ins.2014.04.015

Cavalcante, Y.L., Hauser-Davis, R.A., Saraiva, A.C.F., Brandão, I.L.S., Oliveira, T.F., Silveira, A.M., 2013. Metal and physico-chemical variations at a hydroelectric reservoir analyzed by Multivariate Analyses and Artificial Neural Networks: Environmental management and policy/decision-making tools. Science of The Total Environment 442, 509–514. https://doi.org/10.1016/j.scitotenv.2012.10.059

Chang, F.-J., Chung, C.-H., Chen, P.-A., Liu, C.-W., Coynel, A., Vachaud, G., 2014. Assessment of arsenic concentration in stream water using neuro fuzzy networks with factor analysis. Science of The Total Environment 494–495, 202–210. https://doi.org/10.1016/j.scitotenv.2014.06.133

Chaudhari, S., Patil, M., 2014. Study and Review of Fuzzy Inference Systems for Decision Making and Control.

Chekli, L., Zhao, Y.X., Tijing, L.D., Phuntsho, S., Donner, E., Lombi, E., Gao, B.Y., Shon, H.K., 2015. Aggregation behaviour of engineered nanoparticles in natural waters: Characterising aggregate structure using on-line laser light

scattering. Journal of Hazardous Materials 284, 190–200. https://doi.org/10.1016/j.jhazmat.2014.11.003

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. Environment International 130, 104934. https://doi.org/10.1016/j.envint.2019.104934

Chen, L.Q., Fang, L., Ling, J., Ding, C.Z., Kang, B., Huang, C.Z., 2015. Nanotoxicity of silver nanoparticles to red blood cells: size dependent adsorption, uptake, and hemolytic activity. Chemical research in toxicology 28, 501–509.

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 0.4-2 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. Aquatic Toxicology 174, 46–53. https://doi.org/10.1016/j.aquatox.2016.02.012

Chen, Z., Cao, F., Hu, J., 2015. Approximation by network operators with logistic activation functions. Applied Mathematics and Computation 256, 565–571.

Chiu, S.L., 1996. Selecting Input Variables for Fuzzy Models. Journal of Intelligent and Fuzzy Systems 4, 243–256. https://doi.org/10.3233/IFS-1996-4401

Chiu, S.L., 1994. Fuzzy model identification based on cluster estimation. Journal of Intelligent & fuzzy systems 2, 267–278.

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

Choi, S., Johnston, M., Wang, G.-S., Huang, C.P., 2018. A seasonal observation on the distribution of engineered nanoparticles in municipal wastewater treatment systems exemplified by TiO2 and ZnO. Science of The Total Environment 625, 1321–1329. https://doi.org/10.1016/j.scitotenv.2017.12.326

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. Science of The Total Environment 615, 272–281. https://doi.org/10.1016/j.scitotenv.2017.09.293

Chowdhury, I., Cwiertny, D.M., Walker, S.L., 2012a. 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

Chowdhury, I., Cwiertny, D.M., Walker, S.L., 2012b. 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

Chowdhury, I., Walker, S.L., Mylon, S.E., 2013. Aggregate morphology of nano-TiO $_2$: role of primary particle size, solution chemistry, and organic matter. Environ. Sci.: Processes Impacts 15, 275–282. https://doi.org/10.1039/C2EM30680H

Church, R.M., 2002. The Effective Use of Secondary Data. Learning and Motivation 33, 32–45. https://doi.org/10.1006/lmot.2001.1098

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

Clevert, D.-A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289.

Coll, C., Notter, D., Gottschalk, F., Sun, T., Som, C., Nowack, B., 2016. Probabilistic environmental risk assessment of five nanomaterials (nano-TiO $_2$, nano-Ag, nano-ZnO, CNT, and fullerenes). Nanotoxicology 10, 436–444. https://doi.org/10.3109/17435390.2015.1073812

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

Cordoba, G.A.C., Tuhovčák, L., Tauš, M., 2014. Using Artificial Neural Network Models to Assess Water Quality in Water Distribution Networks. Procedia Engineering 70, 399–408. https://doi.org/10.1016/j.proeng.2014.02.045

Cormier, S.M., Suter, G.W., Zheng, L., 2013. Derivation of a benchmark for freshwater ionic strength. Environmental Toxicology and Chemistry 32, 263–271.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.

Coutinho, F.H., Thompson, C.C., Cabral, A.S., Paranhos, R., Dutilh, B.E., Thompson, F.L., 2019. Modelling the influence of environmental parameters over marine planktonic microbial communities using artificial neural networks. Science of The Total Environment 677, 205–214. https://doi.org/10.1016/j.scitotenv.2019.04.009

Cowan, C.E., 1995. The Multi-media Fate Model: A Vital Tool for Predicting the Fate of Chemicals: Proceeding of a Workshop Organized by the Society of Environmental Toxicology and Chemistry (SETAC): Based on an International Task Force which Addressed the Application of Multi-media Fate Models to Regulatory Decision Making Held at Leuven, Belgium, April 14-16, 1994 and Denver, Colorado, November 4-5, 1994. SETAC press.

Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random forests, in: Ensemble Machine Learning. Springer, pp. 157–175.

D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, T. L. Veith, 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Transactions of the ASABE 50, 885–900. https://doi.org/10.13031/2013.23153

Dagnino, A., Bo, T., Copetta, A., Fenoglio, S., Oliveri, C., Bencivenga, M., Felli, A., Viarengo, A., 2013. Development and application of an innovative expert decision support system to manage sediments and to assess environmental risk in freshwater ecosystems. Environment International 60, 171–182. https://doi.org/10.1016/j.envint.2013.08.011

Dalai, S., Pakrashi, S., Joyce Nirmala, M., Chaudhri, A., Chandrasekaran, N., Mandal, A.B., Mukherjee, A., 2013. Cytotoxicity of TiO2 nanoparticles and their detoxification in a freshwater system. Aquatic Toxicology 138–139, 1–11. https://doi.org/10.1016/j.aquatox.2013.04.005

Dale, A.L., Casman, E.A., Lowry, G.V., Lead, J.R., Viparelli, E., Baalousha, M., 2015. Modeling Nanomaterial Environmental Fate in Aquatic Systems. Environ. Sci. Technol. 49, 2587–2593. https://doi.org/10.1021/es505076w

Danielsson, K., Gallego-Urrea, J.A., Hassellov, M., Gustafsson, S., Jonsson, C.M., 2017. Influence of organic molecules on the aggregation of TiO2 nanoparticles

in acidic conditions. J Nanopart Res 19, 133. https://doi.org/10.1007/s11051-017-3807-9

Danielsson, K., Persson, P., Gallego-Urrea, J.A., Abbas, Z., Rosenqvist, J., Jonsson, C.M., 2018. Effects of the adsorption of NOM model molecules on the aggregation of TiO2 nanoparticles in aqueous suspensions. NanoImpact 10, 177–187. https://doi.org/10.1016/j.impact.2018.05.002

David, C.A., Galceran, J., Rey-Castro, C., Puy, J., Companys, E., Salvador, J., Monné, J., Wallace, R., Vakourov, A., 2012. Dissolution Kinetics and Solubility of ZnO Nanoparticles Followed by AGNES. J. Phys. Chem. C 116, 11758–11767. https://doi.org/10.1021/jp301671b

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. Materials Letters 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. Science of The Total Environment 783, 146984. https://doi.org/10.1016/j.scitotenv.2021.146984

Derjaguin, B.V., 1941. Theory of the Stability of Strongly Charged Lyophobic Sol and of the Adhesion of Strongly Charged Particles in Solutions of Electrolytes. Acta phys. chim. URSS 14, 633.

Dharwal, R., Kaur, L., 2016. Applications of artificial neural networks: a review. Indian J. Sci. Technol 9, 1–8.

Di Addario, M., Temizel, I., Edes, N., Onay, T.T., Demirel, B., Copty, N.K., Ruggeri, B., 2017. Development of Fuzzylogic model to predict the effects of ZnO nanoparticles on methane production from simulated landfill. Journal of Environmental Chemical Engineering 5, 5944–5953. https://doi.org/10.1016/j.jece.2017.10.033

Di Guardo, A., Gouin, T., MacLeod, M., Scheringer, M., 2018. Environmental fate and exposure models: advances and challenges in 21 st century chemical risk assessment. Environmental Science: Processes & Impacts 20, 58–71.

Dietterich, T.G., 1997. Machine-learning research. AI magazine 18, 97–97.

Dogan, E., Sengorur, B., Koklu, R., 2009. Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. Journal of Environmental Management 90, 1229–1235.

Domercq, P., Praetorius, A., Boxall, A.B.A., 2018. Emission and fate modelling framework for engineered nanoparticles in urban aquatic systems at high spatial and temporal resolution. Environ. Sci.: Nano 5, 533–543. https://doi.org/10.1039/C7EN00846E

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

Du, J., Xu, S., Zhou, Q., Li, H., Fu, L., Tang, J., Wang, Y., Peng, X., Xu, Y., Du, X., 2020. A review of microplastics in the aquatic environmental: distribution, transport, ecotoxicology, and toxicological mechanisms. Environmental Science and Pollution Research 27, 11494–11505.

Duan, Y., Coreas, R., Liu, Y., Bitounis, D., Zhang, Z., Parviz, D., Strano, M., Demokritou, P., Zhong, W., 2020. Prediction of protein corona on nanomaterials by machine learning using novel descriptors. NanoImpact 17, 100207. https://doi.org/10.1016/j.impact.2020.100207

Dubois, D., Prade, H., 1980. Systems of linear fuzzy constraints. Fuzzy Sets and Systems 3, 37–48. https://doi.org/10.1016/0165-0114(80)90004-4

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research 12.

Dumont, E., Johnson, A.C., Keller, V.D.J., Williams, R.J., 2015. Nano silver and nano zinc-oxide in surface waters – Exposure estimation for Europe at high spatial and temporal resolution. Environmental Pollution 196, 341–349. https://doi.org/10.1016/j.envpol.2014.10.022

Dwivedi, A.D., Dubey, S.P., Sillanpää, M., Kwon, Y.-N., Lee, C., Varma, R.S., 2015. Fate of engineered nanoparticles: Implications in the environment.

Coordination Chemistry Reviews 287, 64–78. https://doi.org/10.1016/j.ccr.2014.12.014

Ealia, S.A.M., Saravanakumar, M.P., 2017. A review on the classification, characterisation, synthesis of nanoparticles and their application, in: IOP Conference Series: Materials Science and Engineering. IOP Publishing, p. 032019.

El Mahdi, A.M., Aziz, H.A., 2018. A Review on Biodegradation and Toxicity Methods: Risk Assessment, Standards, and Analyses, in: Bidoia, E.D., Montagnolli, R.N. (Eds.), Toxicity and Biodegradation Testing, Methods in Pharmacology and Toxicology. Springer New York, New York, NY, pp. 349–388. https://doi.org/10.1007/978-1-4939-7425-2_18

Ellis, L.-J.A., Valsami-Jones, E., Lead, J.R., Baalousha, M., 2016. Impact of surface coating and environmental conditions on the fate and transport of silver nanoparticles in the aquatic environment. Science of The Total Environment 568, 95–106. https://doi.org/10.1016/j.scitotenv.2016.05.199

Emamgholizadeh, S., Moslemi, K., Karami, G., 2014. Prediction the groundwater level of bastam plain (Iran) by artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS). Water resources management 28, 5433–5446.

Evans, D., Jones, A.J., 2002. A proof of the Gamma test. Proc. R. Soc. Lond. A 458, 2759–2799. https://doi.org/10.1098/rspa.2002.1010

Fang, J., Shijirbaatar, A., Lin, D., Wang, D., Shen, B., Sun, P., Zhou, Z., 2017. Stability of co-existing ZnO and TiO2 nanomaterials in natural water: Aggregation and sedimentation mechanisms. Chemosphere 184, 1125–1133. https://doi.org/10.1016/j.chemosphere.2017.06.097

Fayaz, M., Ullah, I., Park, D.-H., Kim, K., Kim, D., 2017. An Integrated Risk Index Model Based on Hierarchical Fuzzy Logic for Underground Risk Assessment. Applied Sciences 7, 1037. https://doi.org/10.3390/app7101037

Feng, Y., Wang, G., Ruan, L., Du, A., 2017. A finite-volume fast diffusion-limited aggregation model for predicting the coagulation rate of mixed low-ionized system. AIP Advances 7, 035017.

Fernando, I., Zhou, Y., 2019. Impact of pH on the stability, dissolution and aggregation kinetics of silver nanoparticles. Chemosphere 216, 297–305. https://doi.org/10.1016/j.chemosphere.2018.10.122

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

Foldbjerg, R., Dang, D.A., Autrup, H., 2011. Cytotoxicity and genotoxicity of silver nanoparticles in the human lung cancer cell line, A549. Archives of toxicology 85, 743–750.

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. Science of the total environment 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

Franco, A., Trapp, S., 2010. A multimedia activity model for ionizable compounds: Validation study with 2,4-dichlorophenoxyacetic acid, aniline, and trimethoprim. Environ Toxicol Chem 29, 789–799. https://doi.org/10.1002/etc.115

French, R.A., Jacobson, A.R., Kim, B., Isley, S.L., Penn, R.L., Baveye, P.C., 2009. Influence of ionic strength, pH, and cation valence on aggregation kinetics of titanium dioxide nanoparticles. Environmental science & technology 43, 1354–1359.

Frigerio, C., Ribeiro, D.S., Rodrigues, S.S.M., Abreu, V.L., Barbosa, J.A., Prior, J.A., Marques, K.L., Santos, J.L., 2012. Application of quantum dots as analytical tools in automated chemical analysis: a review. Analytica chimica acta 735, 9–22.

Furxhi, I., Murphy, F., Mullins, M., Arvanitis, A., Poland, C.A., 2020. Practices and Trends of Machine Learning Application in Nanotoxicology. Nanomaterials 10, 116. https://doi.org/10.3390/nano10010116

Furxhi, I., Murphy, F., Mullins, M., Poland, C.A., 2019a. Machine learning prediction of nanoparticle in vitro toxicity: A comparative study of classifiers and ensemble-

classifiers using the Copeland Index. Toxicology Letters 312, 157–166. https://doi.org/10.1016/j.toxlet.2019.05.016

Furxhi, I., Murphy, F., Poland, C.A., Sheehan, B., Mullins, M., Mantecca, P., 2019b. Application of Bayesian networks in determining nanoparticle-induced cellular outcomes using transcriptomics. Nanotoxicology 13, 827–848. https://doi.org/10.1080/17435390.2019.1595206

Gacto, M.J., Alcalá, R., Herrera, F., 2011. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. Information Sciences 181, 4340–4360. https://doi.org/10.1016/j.ins.2011.02.021

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

García-Diéguez, C., Herva, M., Roca, E., 2015. A decision support system based on fuzzy reasoning and AHP–FPP for the ecodesign of products: Application to footwear as case study. Applied Soft Computing 26, 224–234. https://doi.org/10.1016/j.asoc.2014.09.043

García-Laencina, P.J., Sancho-Gómez, J.-L., Figueiras-Vidal, A.R., Verleysen, M., 2009. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing 72, 1483–1493. https://doi.org/10.1016/j.neucom.2008.11.026

Gennari, J.H., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubézy, M., Eriksson, H., Noy, N.F., Tu, S.W., 2003. The evolution of Protégé: an environment for knowledge-based systems development. International Journal of Human-computer studies 58, 89–123.

Ghasemi, N., Rohani, S., 2019. Optimization of cyanide removal from wastewaters using a new nano-adsorbent containing ZnO nanoparticles and MOF/Cu and evaluating its efficacy and prediction of experimental results with artificial neural networks. Journal of Molecular Liquids 285, 252–269. https://doi.org/10.1016/j.molliq.2019.04.085

Giese, B., Klaessig, F., Park, B., Kaegi, R., Steinfeldt, M., Wigger, H., von Gleich, A., Gottschalk, F., 2018. Risks, release and concentrations of engineered nanomaterial in the environment. Scientific reports 8, 1565.

Giubilato, E., Zabeo, A., Critto, A., Giove, S., Bierkens, J., Den Hond, E., Marcomini, A., 2014. A risk-based methodology for ranking environmental chemical stressors at the regional scale. Environment International 65, 41–53. https://doi.org/10.1016/j.envint.2013.12.013

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., 2015a. 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

Goldberg, E., Scheringer, M., Bucheli, T.D., Hungerbühler, K., 2015b. 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

Gong, Y., Zhang, Y., Lan, S., Wang, H., 2016. A comparative study of artificial neural networks, support vector machines and adaptive neuro fuzzy inference system for forecasting groundwater levels near Lake Okeechobee, Florida. Water resources management 30, 375–391.

Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y., 2013. Maxout networks, in: International Conference on Machine Learning. PMLR, pp. 1319–1327.

Goodner, K.L., Dreher, J.G., Rouseff, R.L., 2001. The dangers of creating false classifications due to noise in electronic nose and similar multivariate analyses. Sensors and Actuators B: Chemical 80, 261–266.

Gottschalk, F., Kost, E., Nowack, B., 2013. Engineered nanomaterials in water and soils: A risk quantification based on probabilistic exposure and effect modeling: Engineered nanomaterials in water and soils. Environ Toxicol Chem 32, 1278–1287. https://doi.org/10.1002/etc.2177

Gottschalk, F., Lassen, C., Kjoelholt, J., Christensen, F., Nowack, B., 2015. Modeling flows and concentrations of nine engineered nanomaterials in the Danish environment. International journal of environmental research and public health 12, 5581–5602.

Granata, F., Gargano, R., De Marinis, G., 2016. Support vector regression for rainfall-runoff modeling in urban drainage: A comparison with the EPA's storm water management model. Water 8, 69.

Grande, F., Tucci, P., 2016. Titanium dioxide nanoparticles: a risk for human health? Mini reviews in medicinal chemistry 16, 762–769.

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

Grella, T.C., Soares-Lima, H.M., Malaspina, O., Cornélio Ferreira Nocelli, R., 2019. Semi-quantitative analysis of morphological changes in bee tissues: A toxicological approach. Chemosphere 236, 124255. https://doi.org/10.1016/j.chemosphere.2019.06.225

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012. A kernel two-sample test. The Journal of Machine Learning Research 13, 723–773.

Grillo, R., De Jesus, M.B., Fraceto, L.F., 2018. Environmental impact of nanotechnology: analyzing the present for building the future. Frontiers in Environmental Science.

Grillo, R., Rosa, A.H., Fraceto, L.F., 2015. Engineered nanoparticles and organic matter: A review of the state-of-the-art. Chemosphere 119, 608–619. https://doi.org/10.1016/j.chemosphere.2014.07.049

Guan, R., Kang, T., Lu, F., Zhang, Z., Shen, H., Liu, M., 2012. Cytotoxicity, oxidative stress, and genotoxicity in human hepatocyte and embryonic kidney cells exposed to ZnO nanoparticles. Nanoscale research letters 7, 1–7.

Gunsolus, I.L., Mousavi, M.P.S., Hussein, K., Bühlmann, P., Haynes, C.L., 2015. Effects of Humic and Fulvic Acids on Silver Nanoparticle Stability, Dissolution, and Toxicity. Environ. Sci. Technol. 49, 8078–8086. https://doi.org/10.1021/acs.est.5b01496

Guo, Z., Zeng, G., Cui, K., Chen, A., 2019. Toxicity of environmental nanosilver: mechanism and assessment. Environmental Chemistry Letters 17, 319–333.

Gurevitch, J., 1993. Meta-analysis: combining the results of independent experiments. Design and analysis of ecological experiments.

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

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of machine learning research 3, 1157–1182.

Haidich, A.-B., 2010. Meta-analysis in medical research. Hippokratia 14, 29.

Hameed, A.A., Karlik, B., Salman, M.S., 2016. Back-propagation algorithm with variable adaptive momentum. Knowledge-Based Systems 114, 79–87. https://doi.org/10.1016/j.knosys.2016.10.001

Hamilton, S.H., Fu, B., Guillaume, J.H.A., Badham, J., Elsawah, S., Gober, P., Hunt, R.J., Iwanaga, T., Jakeman, A.J., Ames, D.P., Curtis, A., Hill, M.C., Pierce, S.A., Zare, F., 2019. A framework for characterising and evaluating the effectiveness of environmental modelling. Environmental Modelling & Software 118, 83–98. https://doi.org/10.1016/j.envsoft.2019.04.008

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 Research 90, 247–257. https://doi.org/10.1016/j.watres.2015.12.009

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 and Surfaces A: Physicochemical and Engineering Aspects 451, 7–15. https://doi.org/10.1016/j.colsurfa.2014.03.030

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

Hartmann, N.B., Von der Kammer, F., Hofmann, T., Baalousha, M., Ottofuelling, S., Baun, A., 2010. Algal testing of titanium dioxide nanoparticles—testing considerations, inhibitory effects and modification of cadmium bioavailability. Toxicology 269, 190–197.

Hartmann, N.I.B., Skjolding, L.M., Hansen, S.F., Baun, A., Kjølholt, J., Gottschalk, F., 2014. Environmental fate and behaviour of nanomaterials: new knowledge on important transfomation processes.

Haykin, S., Network, N., 2004. A comprehensive foundation. Neural networks 2, 41.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034.

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. Environmental science & technology 53, 4030–4044.

Heinlaan, M., Muna, M., Knöbel, M., Kistler, D., Odzak, N., Kühnel, D., Müller, J., Gupta, G.S., Kumar, A., Shanker, R., Sigg, L., 2016. Natural water as the test medium for Ag and CuO nanoparticle hazard evaluation: An interlaboratory case study. Environmental Pollution 216, 689–699. https://doi.org/10.1016/j.envpol.2016.06.033

Hellendoorn, H., Thomas, C., 1993. Defuzzification in fuzzy controllers. Journal of Intelligent & Fuzzy Systems 1, 109–123.

Hintze, J.L., Nelson, R.D., 1998. Violin Plots: A Box Plot-Density Trace Synergism. The American Statistician 52, 181–184. https://doi.org/10.1080/00031305.1998.10480559

Hochella, M.F., Mogk, D.W., Ranville, J., Allen, I.C., Luther, G.W., Marr, L.C., McGrail, B.P., Murayama, M., Qafoku, N.P., Rosso, K.M., Sahai, N., Schroeder, P.A., Vikesland, P., Westerhoff, P., Yang, Y., 2019. Natural, incidental, and engineered nanomaterials and their impacts on the Earth system. Science 363, eaau8299. https://doi.org/10.1126/science.aau8299

Holden, P.A., Gardea-Torresdey, J.L., Klaessig, F., Turco, R.F., Mortimer, M., Hund-Rinke, K., Cohen Hubal, E.A., Avery, D., Barceló, D., Behra, R., Cohen, Y., Deydier-Stephan, L., Ferguson, P.L., Fernandes, T.F., Herr Harthorn, B., Henderson, W.M., Hoke, R.A., Hristozov, D., Johnston, J.M., Kane, A.B., Kapustka, L., Keller, A.A., Lenihan, H.S., Lovell, W., Murphy, C.J., Nisbet, R.M., Petersen, E.J., Salinas, E.R., Scheringer, M., Sharma, M., Speed, D.E., Sultan, Y., Westerhoff, P., White, J.C., Wiesner, M.R., Wong, E.M., Xing, B., Steele Horan, M., Godwin, H.A., Nel, A.E., 2016. Considerations of Environmentally Relevant Test Conditions for Improved Evaluation of Ecological Hazards of Engineered Nanomaterials. Environ. Sci. Technol. 50, 6124–6145. https://doi.org/10.1021/acs.est.6b00608

Home | Nanotechnology Products Database | NPD [WWW Document], n.d. URL https://product.statnano.com// (accessed 4.23.24).

Hong, E., Yeneneh, A.M., Sen, T.K., Ang, H.M., Kayaalp, A., 2018. ANFIS based Modelling of dewatering performance and polymer dose optimization in a

wastewater treatment plant. Journal of Environmental Chemical Engineering 6, 1957–1968. https://doi.org/10.1016/j.jece.2018.02.041

Hong, R., Pan, T., Qian, J., Li, H., 2006. Synthesis and surface modification of ZnO nanoparticles. Chemical Engineering Journal 119, 71–81. https://doi.org/10.1016/j.cej.2006.03.003

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

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

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

Hristozov, D., Gottardo, S., Semenzin, E., Oomen, A., Bos, P., Peijnenburg, W., van Tongeren, M., Nowack, B., Hunt, N., Brunelli, A., Scott-Fordsmand, J.J., Tran, L., Marcomini, A., 2016. Frameworks and tools for risk assessment of manufactured nanomaterials. Environment International 95, 36–53. https://doi.org/10.1016/j.envint.2016.07.016

Hsiung, C.-E., Lien, H.-L., Galliano, A.E., Yeh, C.-S., Shih, Y., 2016. Effects of water chemistry on the destabilization and sedimentation of commercial TiO2 nanoparticles: Role of double-layer compression and charge neutralization. Chemosphere 151, 145–151. https://doi.org/10.1016/j.chemosphere.2016.02.046

Hu, J., Wang, J., Liu, S., Zhang, Z., Zhang, H., Cai, X., Pan, J., Liu, J., 2018. Effect of TiO2 nanoparticle aggregation on marine microalgae Isochrysis galbana. Journal of environmental sciences 66, 208–215.

Huang, C.-C., Aronstam, R.S., Chen, D.-R., Huang, Y.-W., 2010. Oxidative stress, calcium homeostasis, and altered gene expression in human lung epithelial cells exposed to ZnO nanoparticles. Toxicology in vitro 24, 45–55.

Huang, L., Jolliet, O., 2016. A parsimonious model for the release of volatile organic compounds (VOCs) encapsulated in products. Atmospheric Environment 127, 223–235.

Huynh, K.A., Chen, K.L., 2011. Aggregation Kinetics of Citrate and Polyvinylpyrrolidone Coated Silver Nanoparticles in Monovalent and Divalent Electrolyte Solutions. Environ. Sci. Technol. 45, 5564–5571. https://doi.org/10.1021/es200157h

Jacobs, R., Meesters, J.A.J., ter Braak, C.J.F., van de Meent, D., van der Voet, H., 2016. Combining exposure and effect modeling into an integrated probabilistic environmental risk assessment for nanoparticles: Integrated probabilistic risk assessment for nanoparticles. Environ Toxicol Chem 35, 2958–2967. https://doi.org/10.1002/etc.3476

Jahan, I., Matpan Bekler, F., Tunç, A., Güven, K., 2024. The Effects of Silver Nanoparticles (AgNPs) on Thermophilic Bacteria: Antibacterial, Morphological, Physiological and Biochemical Investigations. Microorganisms 12, 402. https://doi.org/10.3390/microorganisms12020402

Jana, D.K., 2016. Novel arithmetic operations on type-2 intuitionistic fuzzy and its applications to transportation problem. Pacific Science Review A: Natural Science and Engineering 18, 178–189. https://doi.org/10.1016/j.psra.2016.09.008

Jang, J.-S., Sun, C.-T., 1995. Neuro-fuzzy modeling and control. Proceedings of the IEEE 83, 378–406.

Jang, J.-S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst., Man, Cybern. 23, 665–685. https://doi.org/10.1109/21.256541

Jang, J.-S.R., Sun, C.-T., Mizutani, E., 1997. Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book Review]. IEEE Transactions on automatic control 42, 1482–1484.

Janitza, S., Tutz, G., Boulesteix, A.-L., 2016. Random forest for ordinal responses: Prediction and variable selection. Computational Statistics & Data Analysis 96, 57–73. https://doi.org/10.1016/j.csda.2015.10.005

Jansson, N.F., Allen, R.L., Skogsmo, G., Tavakoli, S., 2022. Principal component analysis and K-means clustering as tools during exploration for Zn skarn deposits and industrial carbonates, Sala area, Sweden. Journal of Geochemical Exploration 233, 106909.

Jayalath, S., Wu, H., Larsen, S.C., Grassian, V.H., 2018. Surface Adsorption of Suwannee River Humic Acid on TiO 2 Nanoparticles: A Study of pH and Particle Size. Langmuir 34, 3136–3145. https://doi.org/10.1021/acs.langmuir.8b00300

Jelicic Kadic, A., Vucic, K., Dosenovic, S., Sapunar, D., Puljak, L., 2016. Extracting data from figures with software was faster, with higher interrater reliability than manual extraction. Journal of Clinical Epidemiology 74, 119–123. https://doi.org/10.1016/j.jclinepi.2016.01.002

Jiang, C., Aiken, G.R., Hsu-Kim, H., 2015a. 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

Jiang, C., Aiken, G.R., Hsu-Kim, H., 2015b. 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

Jiang, X., Tong, M., Kim, H., 2012. Influence of natural organic matter on the transport and deposition of zinc oxide nanoparticles in saturated porous media. Journal of Colloid and Interface Science 386, 34–43. https://doi.org/10.1016/j.jcis.2012.07.002

Jimeno-Sáez, P., Senent-Aparicio, J., Pérez-Sánchez, J., Pulido-Velazquez, D., 2018. A comparison of SWAT and ANN models for daily runoff simulation in different climatic zones of peninsular Spain. Water 10, 192.

Johnpaul, C.I., Prasad, M.V., Nickolas, S., Gangadharan, G.R., 2021. Fuzzy representational structures for trend based analysis of time series clustering and classification. Knowledge-Based Systems 222, 106991.

Johnson, A.C., Bowes, M.J., Crossley, A., Jarvie, H.P., Jurkschat, K., Jürgens, M.D., Lawlor, A.J., Park, B., Rowland, P., Spurgeon, D., Svendsen, C., Thompson, I.P., Barnes, R.J., Williams, R.J., Xu, N., 2011. An assessment of the fate, behaviour and environmental risk associated with sunscreen TiO2 nanoparticles in UK field scenarios. Science of The Total Environment 409, 2503–2510. https://doi.org/10.1016/j.scitotenv.2011.03.040

Jones, E.H., Su, C., 2014. Transport and retention of zinc oxide nanoparticles in porous media: Effects of natural organic matter versus natural organic ligands at circumneutral pH. Journal of Hazardous Materials 275, 79–88. https://doi.org/10.1016/j.jhazmat.2014.04.058

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

Kayadelen, C., Taşkıran, T., Günaydın, O., Fener, M., 2009. Adaptive neuro-fuzzy modeling for the swelling potential of compacted soils. Environ Earth Sci 59, 109–115. https://doi.org/10.1007/s12665-009-0009-5

Keller, A.A., McFerran, S., Lazareva, A., Suh, S., 2013. Global life cycle releases of engineered nanomaterials. J Nanopart Res 15, 1692. https://doi.org/10.1007/s11051-013-1692-4

Keller, A.A., Wang, H., Zhou, D., Lenihan, H.S., Cherr, G., Cardinale, B.J., Miller, R., Ji, Z., 2010. Stability and Aggregation of Metal Oxide Nanoparticles in Natural Aqueous Matrices. Environ. Sci. Technol. 44, 1962–1967. https://doi.org/10.1021/es902987d

Kennedy, A.J., Hull, M.S., Bednar, A.J., Goss, J.D., Gunter, J.C., Bouldin, J.L., Vikesland, P.J., Steevens, J.A., 2010. Fractionating nanosilver: importance for determining toxicity to aquatic test organisms. Environmental science & technology 44, 9571–9577.

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

Khan, R., Inam, M., Khan, S., Park, D., Yeom, I., 2019. Interaction between Persistent Organic Pollutants and ZnO NPs in Synthetic and Natural Waters. Nanomaterials 9, 472. https://doi.org/10.3390/nano9030472

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

Kim, I.-S., Baek, M., Choi, S.-J., 2010. Comparative cytotoxicity of Al2O3, CeO2, TiO2 and ZnO nanoparticles to human lung cells. Journal of Nanoscience and Nanotechnology 10, 3453–3458.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Klaine, S.J., Alvarez, P.J.J., Batley, G.E., Fernandes, T.F., Handy, R.D., Lyon, D.Y., Mahendra, S., McLaughlin, M.J., Lead, J.R., 2008. NANOMATERIALS IN THE ENVIRONMENT: BEHAVIOR, FATE, BIOAVAILABILITY, AND EFFECTS. Environ Toxicol Chem 27, 1825. https://doi.org/10.1897/08-090.1

Klasmeier, Jö., Matthies, M., Macleod, M., Fenner, K., Scheringer, M., Stroebe, M., Le Gall, A.C., Mckone, T., Van De Meent, D., Wania, F., 2006. Application of Multimedia Models for Screening Assessment of Long-Range Transport Potential and Overall Persistence. Environ. Sci. Technol. 40, 53–60. https://doi.org/10.1021/es0512024

Klein, J.J.M. de, Quik, J.T.K., Bäuerlein, P.S., Koelmans, A.A., 2016. Towards validation of the NanoDUFLOW nanoparticle fate model for the river Dommel, The Netherlands. Environ. Sci.: Nano 3, 434–441. https://doi.org/10.1039/C5EN00270B

Kovalishyn, V., Abramenko, N., Kopernyk, I., Charochkina, L., Metelytsia, L., Tetko, I.V., Peijnenburg, W., Kustov, L., 2018. Modelling the toxicity of a large set of metal and metal oxide nanoparticles using the OCHEM platform. Food and Chemical Toxicology 112, 507–517. https://doi.org/10.1016/j.fct.2017.08.008

Kozleski, E.B., 2017. The Uses of Qualitative Research: Powerful Methods to Inform Evidence-Based Practice in Education. Research and Practice for Persons with Severe Disabilities 42, 19–32. https://doi.org/10.1177/1540796916683710

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097–1105.

Kumar, M., Prasad Yadav, S., Kumar, S., 2013. Fuzzy system reliability evaluation using time-dependent intuitionistic fuzzy set. International Journal of Systems Science 44, 50–66. https://doi.org/10.1080/00207721.2011.581393

Kumar, P.S., Praveen, T.V., Prasad, M.A., 2016. Artificial neural network model for rainfall-runoff-A case study. International Journal of Hybrid Information Technology 9, 263–272.

Labouta, H.I., Asgarian, N., Rinker, K., Cramb, D.T., 2019. Meta-Analysis of Nanoparticle Cytotoxicity via Data-Mining the Literature. ACS Nano acsnano.8b07562. https://doi.org/10.1021/acsnano.8b07562

Lagos-Avid, M.P., Bonilla, C.A., 2017. Predicting the particle size distribution of eroded sediment using artificial neural networks. Science of The Total Environment 581–582, 833–839. https://doi.org/10.1016/j.scitotenv.2017.01.020

Lai, R.W., Yeung, K.W., Yung, M.M., Djurišić, A.B., Giesy, J.P., Leung, K.M., 2018. Regulation of engineered nanomaterials: current challenges, insights and future directions. Environmental Science and Pollution Research 25, 3060–3077.

Larsen, P., Dai, Y., Collart, F.R., 2015. Predicting bacterial community assemblages using an artificial neural network approach, in: Artificial Neural Networks. Springer, pp. 33–43.

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

Leng, G., McGinnity, T.M., Prasad, G., 2005. An approach for on-line extraction of fuzzy rules using a self-organising fuzzy neural network. Fuzzy sets and systems 150, 211–243.

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. Ecotoxicology and Environmental Safety 243, 113955. https://doi.org/10.1016/j.ecoenv.2022.113955

Li, L., Sillanpää, M., Risto, M., 2016. Influences of water properties on the aggregation and deposition of engineered titanium dioxide nanoparticles in natural waters. Environmental Pollution 219, 132–138. https://doi.org/10.1016/j.envpol.2016.09.080

Li, S., Ma, H., Wallis, L.K., Etterson, M.A., Riley, B., Hoff, D.J., Diamond, S.A., 2016. Impact of natural organic matter on particle behavior and phototoxicity of titanium dioxide nanoparticles. Science of The Total Environment 542, 324–333. https://doi.org/10.1016/j.scitotenv.2015.09.141

Li, S., Sun, W., 2011. A comparative study on aggregation/sedimentation of TiO2 nanoparticles in mono- and binary systems of fulvic acids and Fe(III). Journal

of Hazardous Materials 197, 70–79. https://doi.org/10.1016/j.jhazmat.2011.09.059

Li, X., Lenhart, J.J., Walker, H.W., 2012. Aggregation Kinetics and Dissolution of Coated Silver Nanoparticles. Langmuir 28, 1095–1104. https://doi.org/10.1021/la202328n

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest 2, 6.

Lim, C.C., Kim, H., Vilcassim, M.J.R., Thurston, G.D., Gordon, T., Chen, L.-C., Lee, K., Heimbinder, M., Kim, S.-Y., 2019. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. Environment International 131, 105022. https://doi.org/10.1016/j.envint.2019.105022

Lin, D., Ji, J., Long, Z., Yang, K., Wu, F., 2012. The influence of dissolved and surface-bound humic acid on the toxicity of TiO2 nanoparticles to Chlorella sp. Water research 46, 4477–4487.

Lin, W.-C., Tsai, C.-F., 2020. Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review 53, 1487–1509.

Ling, W., Dong-Mei, F., 2009. Estimation of missing values using a weighted k-nearest neighbors algorithm, in: 2009 International Conference on Environmental Science and Information Application Technology. IEEE, pp. 660–663.

Lipton, Z.C., Berkowitz, J., Elkan, C., 2015. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019.

Liu, H., Wang, Xinxin, Wu, Y., Hou, J., Zhang, S., Zhou, N., Wang, Xiangke, 2019. Toxicity responses of different organs of zebrafish (Danio rerio) to silver nanoparticles with different particle sizes and surface coatings. Environmental Pollution 246, 414–422. https://doi.org/10.1016/j.envpol.2018.12.034

Liu, L., Zhang, Z., Cao, L., Xiong, Z., Tang, Y., Pan, Y., 2021. Cytotoxicity of phytosynthesized silver nanoparticles: A meta-analysis by machine learning algorithms. Sustainable Chemistry and Pharmacy 21, 100425.

Liu, N., Jin, X., Feng, C., Wang, Z., Wu, F., Johnson, A.C., Xiao, H., Hollert, H., Giesy, J.P., 2020. Ecological risk assessment of fifty pharmaceuticals and personal care products (PPCPs) in Chinese surface waters: A proposed multiple-level system. Environment International 136, 105454. https://doi.org/10.1016/j.envint.2019.105454

Liu, W., Sun, W., Borthwick, A.G.L., Ni, J., 2013. Comparison on aggregation and sedimentation of titanium dioxide, titanate nanotubes and titanate nanotubes-TiO2: Influence of pH, ionic strength and natural organic matter. Colloids and Surfaces A: Physicochemical and Engineering Aspects 434, 319–328. https://doi.org/10.1016/j.colsurfa.2013.05.010

Liu, X., Chen, G., Su, C., 2011. Effects of material properties on sedimentation and aggregation of titanium dioxide nanoparticles of anatase and rutile in the aqueous phase. Journal of Colloid and Interface Science 363, 84–91. https://doi.org/10.1016/j.jcis.2011.06.085

Liu, Z., Lu, Y., Shi, Y., Wang, P., Jones, K., Sweetman, A.J., Johnson, A.C., Zhang, M., Zhou, Y., Lu, X., 2017. Crop bioaccumulation and human exposure of perfluoroalkyl acids through multi-media transport from a mega fluorochemical industrial park, China. Environment international 106, 37–47.

Liu, Z., Wang, C., Hou, J., Wang, P., Miao, L., Lv, B., Yang, Y., You, G., Xu, Y., Zhang, M., Ci, H., 2018. Aggregation, sedimentation, and dissolution of CuO and ZnO nanoparticles in five waters. Environ Sci Pollut Res 25, 31240–31249. https://doi.org/10.1007/s11356-018-3123-7

Lo, A., Chernoff, H., Zheng, T., Lo, S.-H., 2015. Why significant variables aren't automatically good predictors. Proc Natl Acad Sci USA 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. Science of The Total Environment 539, 7–16. https://doi.org/10.1016/j.scitotenv.2015.08.115

Loosli, F., Le Coustumer, P., Stoll, S., 2015. Effect of electrolyte valency, alginate concentration and pH on engineered TiO2 nanoparticle stability in aqueous solution. Science of The Total Environment 535, 28–34. https://doi.org/10.1016/j.scitotenv.2015.02.037

Loosli, F., Le Coustumer, P., Stoll, S., 2014. Effect of natural organic matter on the disagglomeration of manufactured TiO2 nanoparticles. Environ. Sci.: Nano 1, 154. https://doi.org/10.1039/c3en00061c

Loosli, F., Le Coustumer, P., Stoll, S., 2013. TiO2 nanoparticles aggregation and disaggregation in presence of alginate and Suwannee River humic acids. pH

and concentration effects on nanoparticle stability. Water Research 47, 6052–6063. https://doi.org/10.1016/j.watres.2013.07.021

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

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

Loukas, Y.L., 2001. Adaptive Neuro-Fuzzy Inference System: An Instant and Architecture-Free Predictor for Improved QSAR Studies. J. Med. Chem. 44, 2772–2783. https://doi.org/10.1021/jm000226c

Lourakis, M.I., 2005. A brief description of the Levenberg-Marquardt algorithm implemented by levmar. Foundation of Research and Technology 4, 1–6.

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

Lowry, G.V., Hill, R.J., Harper, S., Rawle, A.F., Hendren, C.O., Klaessig, F., Nobbmann, U., Sayre, P., Rumble, J., 2016. Guidance to improve the scientific value of zeta-potential measurements in nanoEHS. Environ. Sci.: Nano 3, 953–965. https://doi.org/10.1039/C6EN00136J

Lu, H., Li, H., Liu, T., Fan, Y., Yuan, Y., Xie, M., Qian, X., 2019. Simulating heavy metal concentrations in an aquatic environment using artificial intelligence models and physicochemical indexes. Science of The Total Environment 694, 133591. https://doi.org/10.1016/j.scitotenv.2019.133591

Lu, L., Goerlandt, F., Banda, O.A.V., Kujala, P., 2022. Developing fuzzy logic strength of evidence index and application in Bayesian networks for system risk management. Expert Systems with Applications 192, 116374. https://doi.org/10.1016/j.eswa.2021.116374

Lukoševičius, M., Jaeger, H., 2009. Reservoir computing approaches to recurrent neural network training. Computer Science Review 3, 127–149. https://doi.org/10.1016/j.cosrev.2009.03.005

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proc. Icml. Citeseer, p. 3.

Macko, P., Palosaari, T., Whelan, M., 2021. Extrapolating from acute to chronic toxicity in vitro. Toxicology in Vitro 76, 105206.

MacLeod, M., Scheringer, M., McKone, T.E., Hungerbuhler, K., 2010. The State of Multimedia Mass-Balance Modeling in Environmental Science and Decision-Making. Environ. Sci. Technol. 44, 8360–8364. https://doi.org/10.1021/es100968w

Mahaye, N., Leareng, S.K., Musee, N., 2021. Cytotoxicity and genotoxicity of coated-gold nanoparticles on freshwater algae Pseudokirchneriella subcapitata. Aquatic Toxicology 236, 105865.

Mahaye, N., Musee, N., 2023. Evaluation of Apical and Molecular Effects of Algae Pseudokirchneriella subcapitata to Cerium Oxide Nanoparticles. Toxics 11, 283.

Mahaye, N., Thwala, M., Cowan, D.A., Musee, N., 2017. Genotoxicity of metal based engineered nanoparticles in aquatic organisms: A review. Mutation Research/Reviews in Mutation Research 773, 134–160. https://doi.org/10.1016/j.mrrev.2017.05.004

Mahl, D., Diendorf, J., Meyer-Zaika, W., Epple, M., 2011. Possibilities and limitations of different analytical methods for the size determination of a bimodal dispersion of metallic nanoparticles. Colloids and Surfaces A: Physicochemical and Engineering Aspects 377, 386–392.

Mahmoudi, N., Orouji, H., Fallah-Mehdipour, E., 2016. Integration of shuffled frog leaping algorithm and support vector regression for prediction of water quality parameters. Water resources management 30, 2195–2211.

Majedi, S.M., Kelly, B.C., Lee, H.K., 2014a. Role of combinatorial environmental factors in the behavior and fate of ZnO nanoparticles in aqueous systems: a multiparametric analysis. Journal of hazardous materials 264, 370–379.

Majedi, S.M., Kelly, B.C., Lee, H.K., 2014b. Combined effects of water temperature and chemistry on the environmental fate and behavior of nanosized zinc oxide. Science of The Total Environment 496, 585–593. https://doi.org/10.1016/j.scitotenv.2014.07.082

Majedi, S.M., Kelly, B.C., Lee, H.K., 2014c. Role of combinatorial environmental factors in the behavior and fate of ZnO nanoparticles in aqueous systems: A

multiparametric analysis. Journal of Hazardous Materials 264, 370–379. https://doi.org/10.1016/j.jhazmat.2013.11.015

Malik, A., Tikhamarine, Y., Souag-Gamane, D., Kisi, O., Pham, Q.B., 2020. Support vector regression optimized by meta-heuristic algorithms for daily streamflow prediction. Stochastic Environmental Research and Risk Assessment 34, 1755–1773.

Mamdani, E.H., 1974. Application of fuzzy algorithms for control of simple dynamic plant, in: Proceedings of the Institution of Electrical Engineers. IET, pp. 1585–1588.

Mamdani, E.H., Assilian, S., 1999. An experiment in linguistic synthesis with a fuzzy logic controller. International journal of human-computer studies 51, 135–147.

Mansfield, E.R., Helms, B.P., 1982. Detecting multicollinearity. The American Statistician 36, 158–160.

Matin, S.S., Farahzadi, L., Makaremi, S., Chelgani, S.C., Sattari, Gh., 2018. Variable selection and prediction of uniaxial compressive strength and modulus of elasticity by random forest. Applied Soft Computing 70, 980–987. https://doi.org/10.1016/j.asoc.2017.06.030

Mazhar, S., Ditta, A., Bulgariu, L., Ahmad, I., Ahmed, M., Nadiri, A.A., 2019. Sequential treatment of paper and pulp industrial wastewater: Prediction of water quality parameters by Mamdani Fuzzy Logic model and phytotoxicity assessment. Chemosphere 227, 256–268. https://doi.org/10.1016/j.chemosphere.2019.04.022

McLaughlin, J., Bonzongo, J.-C.J., 2012. Effects of natural water chemistry on nanosilver behavior and toxicity to Ceriodaphnia dubia and Pseudokirchneriella subcapitata. Environmental Toxicology and Chemistry 31, 168–175.

Medina-Velo, I.A., Peralta-Videa, J.R., Gardea-Torresdey, J.L., 2017. Assessing plant uptake and transport mechanisms of engineered nanomaterials from soil. MRS Bulletin 42, 379–384.

Meesters, J.A., Koelmans, A.A., Quik, J.T., Hendriks, A.J., van de Meent, D., 2014a. Multimedia modeling of engineered nanoparticles with SimpleBox4nano: model definition and evaluation. Environmental science & technology 48, 5726–5736.

Meesters, J.A., Koelmans, A.A., Quik, J.T., Hendriks, A.J., van de Meent, D., 2014b. Multimedia modeling of engineered nanoparticles with SimpleBox4nano: model definition and evaluation. Environmental science & technology 48, 5726–5736.

Meesters, J.A., Veltman, K., Hendriks, A.J., van de Meent, D., 2013. Environmental exposure assessment of engineered nanoparticles: Why REACH needs adjustment: Environmental ENPs and REACH. Integr Environ Assess Manag 9, e15–e26. https://doi.org/10.1002/ieam.1446

Meesters, J.A.J., Peijnenburg, W.J.G.M., Hendriks, A.J., Van de Meent, D., Quik, J.T.K., 2019. A model sensitivity analysis to determine the most important physicochemical properties driving environmental fate and exposure of engineered nanoparticles. Environ. Sci.: Nano 6, 2049–2060. https://doi.org/10.1039/C9EN00117D

Men, H., Li, X., Wang, J., Gao, J., 2007. Applies of neural networks to identify gases based on electronic nose, in: 2007 IEEE International Conference on Control and Automation. IEEE, pp. 2699–2704.

Mendel, J.M., 1995. Fuzzy logic systems for engineering: a tutorial. Proceedings of the IEEE 83, 345–377.

Meng, X., Hand, J.L., Schichtel, B.A., Liu, Y., 2018. Space-time trends of PM2.5 constituents in the conterminous United States estimated by a machine learning approach, 2005–2015. Environment International 121, 1137–1147. https://doi.org/10.1016/j.envint.2018.10.029

Mikolajewicz, N., Komarova, S.V., 2019. Meta-Analytic Methodology for Basic Research: A Practical Guide. Front. Physiol. 10, 203. https://doi.org/10.3389/fphys.2019.00203

Miller, T.H., Gallidabino, M.D., MacRae, J.I., Hogstrand, C., Bury, N.R., Barron, L.P., Snape, J.R., Owen, S.F., 2018. Machine Learning for Environmental Toxicology: A Call for Integration and Innovation. Environ. Sci. Technol. 52, 12953–12955. https://doi.org/10.1021/acs.est.8b05382

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

Mirzakhanov, V.E., 2020. Value of fuzzy logic for data mining and machine learning: A case study. Expert Systems with Applications 162, 113781. https://doi.org/10.1016/j.eswa.2020.113781

Mitrano, D.M., Nowack, B., 2017. The need for a life-cycle based aging paradigm for nanomaterials: importance of real-world test systems to identify realistic particle

transformations. Nanotechnology 28, 072001. https://doi.org/10.1088/1361-6528/28/7/072001

Moghaddamnia, A., Ghafari Gousheh, M., Piri, J., Amin, S., Han, D., 2009. Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. Advances in Water Resources 32, 88–97. https://doi.org/10.1016/j.advwatres.2008.10.005

Mohammadi, B., Linh, N.T.T., Pham, Q.B., Ahmed, A.N., Vojteková, J., Guan, Y., Abba, S.I., El-Shafie, A., 2020. Adaptive neuro-fuzzy inference system coupled with shuffled frog leaping algorithm for predicting river streamflow time series. Hydrological Sciences Journal 65, 1738–1751.

Mohd Omar, F., Abdul Aziz, H., Stoll, S., 2014. Aggregation and disaggregation of ZnO nanoparticles: Influence of pH and adsorption of Suwannee River humic acid. Science of The Total Environment 468–469, 195–201. https://doi.org/10.1016/j.scitotenv.2013.08.044

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., PRISMA Group*, the, 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of internal medicine 151, 264–269.

Mohri, M., Rostamizadeh, A., Talwalkar, A., 2018. Foundations of machine learning. MIT press.

Money, E.S., Barton, L.E., Dawson, J., Reckhow, K.H., Wiesner, M.R., 2014. Validation and sensitivity of the FINE Bayesian network for forecasting aquatic exposure to nano-silver. Science of The Total Environment 473–474, 685–691. https://doi.org/10.1016/j.scitotenv.2013.12.100

Money, E.S., Reckhow, K.H., Wiesner, M.R., 2012. The use of Bayesian networks for nanoparticle risk forecasting: Model formulation and baseline evaluation. Science of The Total Environment 426, 436–445. https://doi.org/10.1016/j.scitotenv.2012.03.064

Monikh, F.A., Praetorius, A., Schmid, A., Kozin, P., Meisterjahn, B., Makarova, E., Hofmann, T., von der Kammer, F., 2018. Scientific rationale for the development of an OECD test guideline on engineered nanomaterial stability. NanoImpact 11, 42–50.

Moore, T.L., Rodriguez-Lorenzo, L., Hirsch, V., Balog, S., Urban, D., Jud, C., Rothen-Rutishauser, B., Lattuada, M., Petri-Fink, A., 2015. Nanoparticle colloidal

stability in cell culture media and impact on cellular interactions. Chemical Society Reviews 44, 6287–6305.

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

Müller, B., Reinhardt, J., Strickland, M.T., 1995. Neural networks: an introduction. Springer Science & Business Media.

Muna, M., Blinova, I., Kahru, A., Vinković Vrček, I., Pem, B., Orupõld, K., Heinlaan, M., 2018. Combined Effects of Test Media and Dietary Algae on the Toxicity of CuO and ZnO Nanoparticles to Freshwater Microcrustaceans Daphnia magna and Heterocypris incongruens: Food for Thought. Nanomaterials 9, 23. https://doi.org/10.3390/nano9010023

Musee, 2018. Comment on “risk assessments show engineered nanomaterials to be of low environmental concern.” Environmental science & technology 52, 6723–6724.

Musee, N., 2018. Environmental risk assessment of triclosan and triclocarban from personal care products in South Africa. Environmental Pollution 242, 827–838. https://doi.org/10.1016/j.envpol.2018.06.106

Musee, N., 2017. A model for screening and prioritizing consumer nanoproduct risks: A case study from South Africa. Environment International 100, 121–131. https://doi.org/10.1016/j.envint.2017.01.002

Musee, N., 2011. Simulated environmental risk estimation of engineered nanomaterials: A case of cosmetics in Johannesburg City. Hum Exp Toxicol 30, 1181–1195. https://doi.org/10.1177/0960327110391387

Musee, N., Aldrich, C., Lorenzen, L., 2008. New methodology for hazardous waste classification using fuzzy set theory. Journal of Hazardous Materials 157, 94–105. https://doi.org/10.1016/j.jhazmat.2007.12.104

Musee, N., Lorenzen, L., Aldrich, C., 2006. An aggregate fuzzy hazardous index for composite wastes. Journal of Hazardous Materials 137, 723–733. https://doi.org/10.1016/j.jhazmat.2006.03.060

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. Journal of Environmental Science and Health, Part A 49, 59–66.

Nagy, G., 1991. Neural networks-then and now. IEEE transactions on neural networks 2, 316–318.

Nanochemicals Market Size Report, 2021-2026 [WWW Document], n.d. URL https://www.industryarc.com/Report/16067/nanochemicals-market.html (accessed 4.23.24).

Narita, K., Matsui, Y., Iwao, K., Kamata, M., Matsushita, T., Shirasaki, N., 2014. Selecting pesticides for inclusion in drinking water quality guidelines on the basis of detection probability and ranking. Environment International 63, 114–120. https://doi.org/10.1016/j.envint.2013.10.019

Neal, C., Jarvie, H., Rowland, P., Lawler, A., Sleep, D., Scholefield, P., 2011. Titanium in UK rural, agricultural and urban/industrial rivers: Geogenic and anthropogenic colloidal/sub-colloidal sources and the significance of within-river retention. Science of the Total Environment 409, 1843–1853.

Ni, J., Wu, G.D., Albenberg, L., Tomov, V.T., 2017. Gut microbiota and IBD: causation or correlation? Nature reviews Gastroenterology & hepatology 14, 573–584.

Nielsen, M.A., 2015. Neural networks and deep learning. Determination press San Francisco, CA.

Noori, R., Karbassi, A.R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M.H., Farokhnia, A., Gousheh, M.G., 2011. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. Journal of Hydrology 401, 177–189. https://doi.org/10.1016/j.jhydrol.2011.02.021

Nowack, B., 2017. Evaluation of environmental exposure models for engineered nanomaterials in a regulatory context. NanoImpact 8, 38–47. https://doi.org/10.1016/j.impact.2017.06.005

Nowack, B., Baalousha, M., Bornhöft, N., Chaudhry, Q., Cornelis, G., Cotterill, J., Gondikas, A., Hassellöv, M., Lead, J., Mitrano, D.M., von der Kammer, F., Wontner-Smith, T., 2015. Progress towards the validation of modeled environmental concentrations of engineered nanomaterials by analytical measurements. Environ. Sci.: Nano 2, 421–428. https://doi.org/10.1039/C5EN00100E

Nthwane, Y.B., Tancu, Y., Maity, A., Thwala, M., 2019. Characterisation of titanium oxide nanomaterials in sunscreens obtained by extraction and release exposure scenarios. SN Applied Sciences 1, 312.

Nyangiwe, N.N., Ouma, C.N.M., 2019. Modelling the adsorption of natural organic matter on Ag (111) surface: Insights from dispersion corrected density functional theory calculations. Journal of Molecular Graphics and Modelling 92, 313–319. https://doi.org/10.1016/j.jmgm.2019.08.013

Obiedat, M., Samarasinghe, S., 2016. A novel semi-quantitative Fuzzy Cognitive Map model for complex systems for addressing challenging participatory real life problems. Applied Soft Computing 48, 91–110. https://doi.org/10.1016/j.asoc.2016.06.001

Odzak, N., Kistler, D., Sigg, L., 2017a. Influence of daylight on the fate of silver and zinc oxide nanoparticles in natural aquatic environments. Environmental Pollution 226, 1–11. https://doi.org/10.1016/j.envpol.2017.04.006

Odzak, N., Kistler, D., Sigg, L., 2017b. Influence of daylight on the fate of silver and zinc oxide nanoparticles in natural aquatic environments. Environmental Pollution 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 transactions on computational biology and bioinformatics 17, 2131–2140.

Ojala, M., Garriga, G.C., 2010. Permutation tests for studying classifier performance. Journal of machine learning research 11.

Olyaie, E., Banejad, H., Chau, K.-W., Melesse, A.M., 2015. A comparison of various artificial intelligence approaches performance for estimating suspended sediment load of river systems: a case study in United States. Environmental monitoring and assessment 187, 1–22.

Omid, M., Lashgari, M., Mobli, H., Alimardani, R., Mohtasebi, S., Hesamifard, R., 2010. Design of fuzzy logic control system incorporating human expert knowledge for combine harvester. Expert Systems with Applications 37, 7080–7085. https://doi.org/10.1016/j.eswa.2010.03.010

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 Engineering Journal 12, 1545–1556.

Ottofuelling, S., Von Der Kammer, F., Hofmann, T., 2011. Commercial Titanium Dioxide Nanoparticles in Both Natural and Synthetic Water: Comprehensive Multidimensional Testing and Prediction of Aggregation Behavior. Environ. Sci. Technol. 45, 10045–10052. https://doi.org/10.1021/es2023225

Palomino, D., Yamunake, C., Coustumer, P.L., Stoll, S., 2013. Stability of TiO₂ Nanoparticles in Presence of Fulvic Acids. Importance of pH. j coll sci biotechnol 2, 62–69. https://doi.org/10.1166/jcsb.2013.1033

Pan, L., Li, J., 2010. K-nearest neighbor based missing data estimation algorithm in wireless sensor networks. Wireless Sensor Network 2, 115.

Pang, W., Coghill, G.M., 2015a. Qualitative, semi-quantitative, and quantitative simulation of the osmoregulation system in yeast. Biosystems 131, 40–50. https://doi.org/10.1016/j.biosystems.2015.04.003

Pang, W., Coghill, G.M., 2015b. Qualitative, semi-quantitative, and quantitative simulation of the osmoregulation system in yeast. Biosystems 131, 40–50. https://doi.org/10.1016/j.biosystems.2015.04.003

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 and QSAR in Environmental Research 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. REVIEWS ON ADVANCED MATERIALS SCIENCE 53, 119–130. https://doi.org/10.1515/rams-2018-0009

Patterson, D.W., 1990. Introduction to artificial intelligence and expert systems. Prentice-hall of India.

Paul, A.K., Shill, P.C., Rabin, Md.R.I., Murase, K., 2018. Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease. Appl Intell 48, 1739–1756. https://doi.org/10.1007/s10489-017-1037-6

Peng, C., Zhang, W., Gao, H., Li, Y., Tong, X., Li, K., Zhu, X., Wang, Y., Chen, Y., 2017. Behavior and Potential Impacts of Metal-Based Engineered

Nanoparticles in Aquatic Environments. Nanomaterials 7, 21. https://doi.org/10.3390/nano7010021

Peng, J., Jury, E.C., Dönnes, P., Ciurtin, C., 2021. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. Frontiers in pharmacology 12, 720694.

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. Environmental Pollution 267, 115434. https://doi.org/10.1016/j.envpol.2020.115434

Peng, Y.-H., Tsai, Y.-C., Hsiung, C.-E., Lin, Y.-H., Shih, Y., 2017a. Influence of water chemistry on the environmental behaviors of commercial ZnO nanoparticles in various water and wastewater samples. Journal of Hazardous Materials 322, 348–356. https://doi.org/10.1016/j.jhazmat.2016.10.003

Peng, Y.-H., Tsai, Y.-C., Hsiung, C.-E., Lin, Y.-H., Shih, Y., 2017b. Influence of water chemistry on the environmental behaviors of commercial ZnO nanoparticles in various water and wastewater samples. Journal of Hazardous Materials 322, 348–356. https://doi.org/10.1016/j.jhazmat.2016.10.003

Peng, Y.-H., Tsai, Y.-C., Hsiung, C.-E., Lin, Y.-H., Shih, Y., 2017c. Influence of water chemistry on the environmental behaviors of commercial ZnO nanoparticles in various water and wastewater samples. Journal of Hazardous Materials 322, 348–356. https://doi.org/10.1016/j.jhazmat.2016.10.003

Peng, Y.-H., Tso, C., Tsai, Y., Zhuang, C., Shih, Y., 2015. The effect of electrolytes on the aggregation kinetics of three different ZnO nanoparticles in water. Science of The Total Environment 530–531, 183–190. https://doi.org/10.1016/j.scitotenv.2015.05.059

Pepa, L., Capecci, M., Andrenelli, E., Ciabattoni, L., Spalazzi, L., Ceravolo, M.G., 2020. A fuzzy logic system for the home assessment of freezing of gait in subjects with Parkinsons disease. Expert Systems with Applications 147, 113197. https://doi.org/10.1016/j.eswa.2020.113197

Peretyazhko, T.S., Zhang, Q., Colvin, V.L., 2014. Size-Controlled Dissolution of Silver Nanoparticles at Neutral and Acidic pH Conditions: Kinetics and Size Changes. Environ. Sci. Technol. 48, 11954–11961. https://doi.org/10.1021/es5023202

Peters, R.J.B., Van Bemmel, G., Milani, N.B.L., Den Hertog, G.C.T., Undas, A.K., Van Der Lee, M., Bouwmeester, H., 2018. Detection of nanoparticles in Dutch

surface waters. Science of The Total Environment 621, 210–218. https://doi.org/10.1016/j.scitotenv.2017.11.238

Pettitt, M.E., Lead, J.R., 2013. Minimum physicochemical characterisation requirements for nanomaterial regulation. Environment International 52, 41–50. https://doi.org/10.1016/j.envint.2012.11.009

Pham, B.T., Nguyen, M.D., Dao, D.V., Prakash, I., Ly, H.-B., Le, T.-T., Ho, L.S., Nguyen, K.T., Ngo, T.Q., Hoang, V., Son, L.H., Ngo, H.T.T., Tran, H.T., Do, N.M., Van Le, H., Ho, H.L., Tien Bui, D., 2019. Development of artificial intelligence models for the prediction of Compression Coefficient of soil: An application of Monte Carlo sensitivity analysis. Science of The Total Environment 679, 172–184. https://doi.org/10.1016/j.scitotenv.2019.05.061

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

Piccinno, F., Gottschalk, F., Seeger, S., Nowack, B., 2012. Industrial production quantities and uses of ten engineered nanomaterials in Europe and the world. J Nanopart Res 14, 1109. https://doi.org/10.1007/s11051-012-1109-9

Pilone, E., Demichela, M., 2018. A semi-quantitative methodology to evaluate the main local territorial risks and their interactions. Land Use Policy 77, 143–154. https://doi.org/10.1016/j.landusepol.2018.05.027

Poborilova, Z., Opatrilova, R., Babula, P., 2013. Toxicity of aluminium oxide nanoparticles demonstrated using a BY-2 plant cell suspension culture model. Environmental and experimental Botany 91, 1–11.

Poul, A.K., Shourian, M., Ebrahimi, H., 2019. A comparative study of MLR, KNN, ANN and ANFIS models with wavelet transform in monthly stream flow prediction. Water Resources Management 33, 2907–2923.

Pradhan, P., Tingsanchali, T., Shrestha, S., 2020. Evaluation of Soil and Water Assessment Tool and Artificial Neural Network models for hydrologic simulation in different climatic regions of Asia. Science of The Total Environment 701, 134308. https://doi.org/10.1016/j.scitotenv.2019.134308

Praetorius, A., Badetti, E., Brunelli, A., Clavier, A., Gallego-Urrea, J.A., Gondikas, A., Hassellöv, M., Hofmann, T., Mackevica, A., Marcomini, A., 2020. Strategies for determining heteroaggregation attachment efficiencies of engineered

nanoparticles in aquatic environments. Environmental Science: Nano 7, 351–367.

Praetorius, A., Labille, J., Scheringer, M., Thill, A., Hungerbühler, K., Bottero, J.-Y., 2014. Heteroaggregation of titanium dioxide nanoparticles with model natural colloids under environmentally relevant conditions. Environmental science & technology 48, 10690–10698.

Pratama, M., Er, M.J., Li, X., Oentaryo, R.J., Lughofer, E., Arifin, I., 2013. Data driven modeling based on dynamic parsimonious fuzzy neural network. Neurocomputing 110, 18–28.

Probst, P., Boulesteix, A.-L., 2017. To tune or not to tune the number of trees in random forest. The Journal of Machine Learning Research 18, 6673–6690.

Pushpa, P., Manimala, K., 2014. Implementation of hyperbolic tangent activation function in VLSI. International Journal of Advanced Research in Computer Science & Technology 2, 225–228.

Qi, J., Ye, Y.Y., Wu, J.J., Wang, H.T., Li, F.T., 2013. Dispersion and stability of titanium dioxide nanoparticles in aqueous suspension: effects of ultrasonication and concentration. Water Science and Technology 67, 147–151. https://doi.org/10.2166/wst.2012.545

Quan, Q., Hao, Z., Xifeng, H., Jingchun, L., 2020. Research on water temperature prediction based on improved support vector regression. Neural Computing and Applications 1–10.

Rahman, M.M., Charoenlarpnopparut, C., Suksompong, P., 2015. Classification and pattern recognition algorithms applied to E-Nose, in: 2015 2nd International Conference on Electrical Information and Communication Technologies (EICT). IEEE, pp. 44–48.

Rajaee, T., Mirbagheri, S.A., Zounemat-Kermani, M., Nourani, V., 2009. Daily suspended sediment concentration simulation using ANN and neuro-fuzzy models. Science of The Total Environment 407, 4916–4927. https://doi.org/10.1016/j.scitotenv.2009.05.016

Rajkovic, S., Bornhöft, N.A., van der Weijden, R., Nowack, B., Adam, V., 2020. Dynamic probabilistic material flow analysis of engineered nanomaterials in European waste treatment systems. Waste Management 113, 118–131.

Ramirez, E.M., Mayorga, R.V., 2008. On the Parameter Optimization of Fuzzy Inference Systems.

Ramirez, R., Martí, V., Darbra, R.M., 2022. Environmental Risk Assessment of Silver Nanoparticles in Aquatic Ecosystems Using Fuzzy Logic. Water 14, 1885.

Ranjan, S., Ramalingam, C., 2016. Titanium dioxide nanoparticles induce bacterial membrane rupture by reactive oxygen species generation. Environmental Chemistry Letters 14, 487–494.

Raza, G., Amjad, M., Kaur, I., Wen, D., 2016. Stability and Aggregation Kinetics of Titania Nanomaterials under Environmentally Realistic Conditions. Environ. Sci. Technol. 50, 8462–8472. https://doi.org/10.1021/acs.est.5b05746

Rehman, M.Z., Nawi, N.M., 2011. The Effect of Adaptive Momentum in Improving the Accuracy of Gradient Descent Back Propagation Algorithm on Classification Problems, in: Mohamad Zain, J., Wan Mohd, W.M. bt, El-Qawasmeh, E. (Eds.), Software Engineering and Computer Systems, Communications in Computer and Information Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 380–390. https://doi.org/10.1007/978-3-642-22170-5_33

Reitermanova, Z., 2010. Data splitting, in: WDS. Matfyzpress Prague, pp. 31–36.

Remesan, R., Mathew, J., 2015. Hydrological Data Driven Modelling: A Case Study Approach. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-09235-5

Rivera-Utrilla, J., Sánchez-Polo, M., Ferro-García, M.Á., Prados-Joya, G., Ocampo-Pérez, R., 2013. Pharmaceuticals as emerging contaminants and their removal from water. A review. Chemosphere 93, 1268–1287.

Robitaille, B., Marcos, B., Veillette, M., Payre, G., 1970. Quasi-Newton methods for training neural networks. WIT Transactions on Information and Communication Technologies 2.

Rojas, R., 1996. The backpropagation algorithm, in: Neural Networks. Springer, pp. 149–182.

Romanello, M.B., Fidalgo de Cortalezzi, M.M., 2013. An experimental study on the aggregation of TiO2 nanoparticles under environmentally relevant conditions. Water Research 47, 3887–3898. https://doi.org/10.1016/j.watres.2012.11.061

Rosati, R., Romeo, L., Cecchini, G., Tonetto, F., Viti, P., Mancini, A., Frontoni, E., 2023. From knowledge-based to big data analytic model: a novel IoT and machine learning based decision support system for predictive maintenance in Industry 4.0. J Intell Manuf 34, 107–121. https://doi.org/10.1007/s10845-022-01960-x

Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65, 386–408. https://doi.org/10.1037/h0042519

Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychological bulletin 86, 638.

Roy, K., Mukherjee, A., Jana, D.K., 2019. Prediction of maximum oil-yield from almond seed in a chemical industry: A novel type-2 fuzzy logic approach. South African Journal of Chemical Engineering 29, 1–9. https://doi.org/10.1016/j.sajce.2019.03.001

Rui, J., Zhang, H., Zhang, D., Han, F., Guo, Q., 2019. Total organic carbon content prediction based on support-vector-regression machine with particle swarm optimization. Journal of Petroleum Science and Engineering 180, 699–706.

Runkler, T.A., 1997. Selection of appropriate defuzzification methods using application specific properties. IEEE Trans. Fuzzy Syst. 5, 72–79. https://doi.org/10.1109/91.554449

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

Saaty, R.W., 1987. The analytic hierarchy process—what it is and how it is used. Mathematical Modelling 9, 161–176. https://doi.org/10.1016/0270-0255(87)90473-8

Saaty, T.L., 2016. The Analytic Hierarchy and Analytic Network Processes for the Measurement of Intangible Criteria and for Decision-Making, in: Greco, S., Ehrgott, M., Figueira, J.R. (Eds.), Multiple Criteria Decision Analysis, International Series in Operations Research & Management Science. Springer New York, New York, NY, pp. 363–419. https://doi.org/10.1007/978-1-4939-3094-4_10

Saaty, T.L., 1980. The analytical hierarchy process, planning, priority. Resource allocation. RWS publications, USA.

Sackett, D.L., Rosenberg, W.M., Gray, J.M., Haynes, R.B., Richardson, W.S., 1996. Evidence based medicine: what it is and what it isn't. British Medical Journal Publishing Group.

Sadollah, A. (Ed.), 2018. Fuzzy Logic Based in Optimization Methods and Control Systems and its Applications. InTech. https://doi.org/10.5772/intechopen.73112

Saeys, Y., Inza, I., Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Sahoo, G.B., Ray, C., Mehnert, E., Keefer, D.A., 2006. Application of artificial neural networks to assess pesticide contamination in shallow groundwater. Science of The Total Environment 367, 234–251. https://doi.org/10.1016/j.scitotenv.2005.12.011

Sani-Kast, N., Scheringer, M., Slomberg, D., Labille, J., Praetorius, A., Ollivier, P., Hungerbühler, K., 2015. Addressing the complexity of water chemistry in environmental fate modeling for engineered nanoparticles. Science of the Total Environment 535, 150–159.

Santos, E.C., Armas, E.D., Crowley, D., Lambais, M.R., 2014. Artificial neural network modeling of microbial community structures in the Atlantic Forest of Brazil. Soil Biology and Biochemistry 69, 101–109.

Sarkar, A., Pandey, P., 2015. River Water Quality Modelling Using Artificial Neural Network Technique. Aquatic Procedia 4, 1070–1077. https://doi.org/10.1016/j.aqpro.2015.02.135

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. Science of The Total Environment 535, 3–19. https://doi.org/10.1016/j.scitotenv.2014.10.035

Schölkopf, B., Smola, A.J., 2002. Learning with kernels: support vector machines, Regularization. Optimization, and Beyond. MIT press 1.

Selck, H., Handy, R.D., Fernandes, T.F., Klaine, S.J., Petersen, E.J., 2016. Nanomaterials in the aquatic environment: A European Union–United States perspective on the status of ecotoxicity testing, research priorities, and challenges ahead. Environ Toxicol Chem 35, 1055–1067. https://doi.org/10.1002/etc.3385

Sengul, A.B., Asmatulu, E., 2020. Toxicity of metal and metal oxide nanoparticles: A review. Environmental Chemistry Letters 18, 1659–1683.

Shandilya, N., Le Bihan, O., Bressot, C., Morgeneyer, M., 2015. Emission of Titanium Dioxide Nanoparticles from Building Materials to the Environment by Wear and Weather. Environ. Sci. Technol. 49, 2163–2170. https://doi.org/10.1021/es504710p

Shao, H., Zheng, G., 2011. Convergence analysis of a back-propagation algorithm with adaptive momentum. Neurocomputing 74, 749–752. https://doi.org/10.1016/j.neucom.2010.10.008

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. Advances in Colloid and Interface Science 204, 15–34. https://doi.org/10.1016/j.cis.2013.12.002

Shi, H., Magaye, R., Castranova, V., Zhao, J., 2013. Titanium dioxide nanoparticles: a review of current toxicological data. Particle and fibre toxicology 10, 1–33.

Shi, Y., Eberhart, R., Chen, Y., 1999. Implementation of evolutionary fuzzy systems. IEEE Transactions on fuzzy systems 7, 109–119.

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. International journal of pharmaceutics 453, 12–24.

Singer, A., Gray, S., Sadler, A., Schmitt Olabisi, L., Metta, K., Wallace, R., Lopez, M.C., Introne, J., Gorman, M., Henderson, J., 2017. Translating community narratives into semi-quantitative models to understand the dynamics of socio-environmental crises. Environmental Modelling & Software 97, 46–55. https://doi.org/10.1016/j.envsoft.2017.07.010

Siregar, D., Arisandi, D., Usman, A., Irwan, D., Rahim, R., 2017. Research of simple multi-attribute rating technique for decision support, in: Journal of Physics: Conference Series. IOP Publishing, p. 012015.

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. Ecotoxicology and Environmental Safety 185, 109733. https://doi.org/10.1016/j.ecoenv.2019.109733

Smola, A.J., Murata, N., Schölkopf, B., Müller, K.-R., 1998. Asymptotically optimal choice of ε-loss for support vector machines, in: International Conference on Artificial Neural Networks. Springer, pp. 105–110.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and computing 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. Environmental Science: Nano 10, 1790–1799.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., de' Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I., Schwartz, J., 2019. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environment International 124, 170–179. https://doi.org/10.1016/j.envint.2019.01.016

Stefánsson, A., Končar, N., Jones, A.J., 1997. A note on the Gamma test. Neural Comput & Applic 5, 131–133. https://doi.org/10.1007/BF01413858

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9, 307. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25. https://doi.org/10.1186/1471-2105-8-25

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. Infection, Genetics and Evolution 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

Sun, T., Zhan, J., Li, F., Ji, C., Wu, H., 2021. Evidence-based meta-analysis of the genotoxicity induced by microplastics in aquatic organisms at environmentally relevant concentrations. Science of The Total Environment 783, 147076. https://doi.org/10.1016/j.scitotenv.2021.147076

Sun, T.Y., Bornhöft, N.A., Hungerbühler, K., Nowack, B., 2016. Dynamic Probabilistic Modeling of Environmental Emissions of Engineered Nanomaterials. Environ. Sci. Technol. 50, 4701–4711. https://doi.org/10.1021/acs.est.5b05828

Sun, T.Y., Gottschalk, F., Hungerbühler, K., Nowack, B., 2014. Comprehensive probabilistic modelling of environmental emissions of engineered nanomaterials. Environmental Pollution 185, 69–76. https://doi.org/10.1016/j.envpol.2013.10.004

Swihart, M.T., 2003. Vapor-phase synthesis of nanoparticles. Current Opinion in Colloid & Interface Science 8, 127–133.

Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its applications to modeling and control. IEEE transactions on systems, man, and cybernetics 116–132.

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

Tan, Y., Shuai, C., Jiao, L., Shen, L., 2017. An adaptive neuro-fuzzy inference system (ANFIS) approach for measuring country sustainability performance. Environmental Impact Assessment Review 65, 29–40. https://doi.org/10.1016/j.eiar.2017.04.004

Tang, C., Tan, J., Fan, Y., Zheng, K., Yu, Z., Peng, X., 2019. Quantitative and semiquantitative analyses of hexa-mix-chlorinated/brominated benzenes in fly ash, soil and air using gas chromatography-high resolution mass spectrometry assisted with isotopologue distribution computation. Environmental Pollution 255, 113162. https://doi.org/10.1016/j.envpol.2019.113162

Tangaa, S.R., Selck, H., Winther-Nielsen, M., Khan, F.R., 2016. Trophic transfer of metal-based nanoparticles in aquatic environments: a review and recommendations for future research focus. Environmental Science: Nano 3, 966–981.

Taniguchi, T., Tanaka, K., Ohtake, H., Wang, H.O., 2001. Model construction, rule reduction, and robust compensation for generalized form of Takagi-Sugeno fuzzy systems. IEEE Trans. Fuzzy Syst. 9, 525–538. https://doi.org/10.1109/91.940966

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. 106, 7183–7192. https://doi.org/10.1029/2000JD900719

Tejamaya, M., Römer, I., Merrifield, R.C., Lead, J.R., 2012. Stability of Citrate, PVP, and PEG Coated Silver Nanoparticles in Ecotoxicology Media. Environ. Sci. Technol. 46, 7011–7017. https://doi.org/10.1021/es2038596

Theodoridis, S., 2015. Chapter 18 - Neural Networks and Deep Learning, in: Theodoridis, S. (Ed.), Machine Learning. Academic Press, Oxford, pp. 875–936. https://doi.org/10.1016/B978-0-12-801522-3.00018-5

Thio, B.J.R., Zhou, D., Keller, A.A., 2011. Influence of natural organic matter on the aggregation and deposition of titanium dioxide nanoparticles. Journal of Hazardous Materials 189, 556–563. https://doi.org/10.1016/j.jhazmat.2011.02.072

Thwala, M., Klaine, S.J., Musee, N., 2016. Interactions of metal-based engineered nanoparticles with aquatic higher plants: A review of the state of current knowledge. Environmental Toxicology and Chemistry 35, 1677–1694.

Tiede, K., Hanssen, S.F., Westerhoff, P., Fern, G.J., Hankin, S.M., Aitken, R.J., Chaudhry, Q., Boxall, A.B.A., 2015. How important is drinking water exposure for the risks of engineered nanoparticles to consumers? Nanotoxicology 1–9. https://doi.org/10.3109/17435390.2015.1022888

Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning 4, 26–31.

Tolaymat, T., El Badawy, A., Genaidy, A., Abdelraheem, W., Sequeira, R., 2017. Analysis of metallic and metal oxide nanomaterial environmental emissions.

Journal of Cleaner Production 143, 401–412. https://doi.org/10.1016/j.jclepro.2016.12.094

Tolaymat, T., El Badawy, A., Sequeira, R., Genaidy, A., 2015. An integrated science-based methodology to assess potential risks and implications of engineered nanomaterials. Journal of Hazardous Materials 298, 270–281. https://doi.org/10.1016/j.jhazmat.2015.04.019

Topuz, E., van Gestel, C.A.M., 2016. An approach for environmental risk assessment of engineered nanomaterials using Analytical Hierarchy Process (AHP) and fuzzy inference rules. Environment International 92–93, 334–347. https://doi.org/10.1016/j.envint.2016.04.022

Traore, A., Grieu, S., Puig, S., Corominas, L., Thiéry, F., Polit, M., Colprim, J., 2005. Fuzzy control of dissolved oxygen in a sequencing batch reactor pilot plant. Chemical Engineering Journal 111, 13–19.

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. Environmental Science: Nano 5, 1902–1910.

Troester, M., Brauch, H.-J., Hofmann, T., 2016. Vulnerability of drinking water supplies to engineered nanoparticles. Water research 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

Tso, C., Zhung, C., Shih, Y., Tseng, Y.-M., Wu, S., Doong, R., 2010. Stability of metal oxide nanoparticles in aqueous solutions. Water Science and Technology 61, 127–133. https://doi.org/10.2166/wst.2010.787

Uskoković, V., 2012. Dynamic light scattering based microelectrophoresis: main prospects and limitations. Journal of dispersion science and technology 33, 1762–1786.

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.

Valizadeh, M., Sohrabi, M.R., 2018. The application of artificial neural networks and support vector regression for simultaneous spectrophotometric determination of commercial eye drop contents. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 193, 297–304. https://doi.org/10.1016/j.saa.2017.11.056

Van den Brink, P.J., Bracewell, S.A., Bush, A., Chariton, A., Choung, C.B., Compson, Z.G., Dafforn, K.A., Korbel, K., Lapen, D.R., Mayer-Pinto, M., Monk, W.A., O'Brien, A.L., Rideout, N.K., Schäfer, R.B., Sumon, K.A., Verdonschot, R.C.M., Baird, D.J., 2019. Towards a general framework for the assessment of interactive effects of multiple stressors on aquatic ecosystems: Results from the Making Aquatic Ecosystems Great Again (MAEGA) workshop. Science of The Total Environment 684, 722–726. https://doi.org/10.1016/j.scitotenv.2019.02.455

Van Der Malsburg, C., 1986. Frank rosenblatt: Principles of neurodynamics: Perceptrons and the theory of brain mechanisms, in: Brain Theory. Springer, pp. 245–248.

Vance, M.E., Kuiken, T., Vejerano, E.P., McGinnis, S.P., Hochella, M.F., Rejeski, D., Hull, M.S., 2015. Nanotechnology in the real world: Redeveloping the nanomaterial consumer products inventory. Beilstein J. Nanotechnol. 6, 1769–1780. https://doi.org/10.3762/bjnano.6.181

Van't Veer, L.J., Dai, H., Van De Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., Van Der Kooy, K., Marton, M.J., Witteveen, A.T., 2002. Gene expression profiling predicts clinical outcome of breast cancer. nature 415, 530–536.

Vapnik, V.N., 1995. T he Nature of Statistical L earning T heory. New York: Springer2V erlag.

Venkatesan, A.K., Reed, R.B., Lee, S., Bi, X., Hanigan, D., Yang, Y., Ranville, J.F., Herckes, P., Westerhoff, P., 2018. Detection and Sizing of Ti-Containing Particles in Recreational Waters Using Single Particle ICP-MS. Bull Environ Contam Toxicol 100, 120–126. https://doi.org/10.1007/s00128-017-2216-1

Verwey, E.J.W., Overbeek, J.T.G., 1955. Theory of the stability of lyophobic colloids. Journal of Colloid Science 10, 224–225.

Vieira, G.C., de Mendonça, A.R., da Silva, G.F., Zanetti, S.S., da Silva, M.M., dos Santos, A.R., 2018. Prognoses of diameter and height of trees of eucalyptus using artificial intelligence. Science of The Total Environment 619–620, 1473–1481. https://doi.org/10.1016/j.scitotenv.2017.11.138

Wagner, S., Gondikas, A., Neubauer, E., Hofmann, T., von der Kammer, F., 2014. Spot the Difference: Engineered and Natural Nanoparticles in the Environment-Release, Behavior, and Fate. Angew. Chem. Int. Ed. n/a-n/a. https://doi.org/10.1002/anie.201405050

Walker, E., Hernandez, A.V., Kattan, M.W., 2008. Meta-analysis: Its strengths and limitations. Cleveland Clinic journal of medicine 75, 431.

Wang, H., Adeleye, A.S., Huang, Y., Li, F., Keller, A.A., 2015. Heteroaggregation of nanoparticles with biocolloids and geocolloids. Advances in colloid and interface Science 226, 24–36.

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. Knowledge-Based Systems 53, 100–107. https://doi.org/10.1016/j.knosys.2013.08.025

Wang, X., Liu, L., Zhang, W., Ma, X., 2021. Prediction of Plant Uptake and Translocation of Engineered Metallic Nanoparticles by Machine Learning. Environ. Sci. Technol. 55, 7491–7500. https://doi.org/10.1021/acs.est.1c01603

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. Environment International 133, 105161. https://doi.org/10.1016/j.envint.2019.105161

Wang, Z., Gu, X., Ouyang, W., Lin, C., Zhu, J., Xu, L., Liu, X., He, M., Wang, B., 2020. Trophodynamics of arsenic for different species in coastal regions of the Northwest Pacific Ocean: In situ evidence and a meta-analysis. Water Research 184, 116186. https://doi.org/10.1016/j.watres.2020.116186

Wang, Z., Luo, Z., Yan, Y., 2018. Dispersion and sedimentation of titanium dioxide nanoparticles in freshwater algae and daphnia aquatic culture media in the

presence of arsenate. Journal of Experimental Nanoscience 13, 119–129. https://doi.org/10.1080/17458080.2018.1449023

Webster, E., Mackay, D., Wania, F., 1998. Evaluating environmental persistence. Environmental Toxicology and Chemistry: An International Journal 17, 2148–2158.

Weir, A., Westerhoff, P., Fabricius, L., Hristovski, K., Von Goetz, N., 2012. Titanium dioxide nanoparticles in food and personal care products. Environmental science & technology 46, 2242–2250.

Welch, I., Goyal, A., 2008. A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. Rev. Financ. Stud. 21, 1455–1508. https://doi.org/10.1093/rfs/hhm014

Williams, R.J., Harrison, S., Keller, V., Kuenen, J., Lofts, S., Praetorius, A., Svendsen, C., Vermeulen, L.C., van Wijnen, J., 2019. Models for assessing engineered nanomaterial fate and behaviour in the aquatic environment. Current Opinion in Environmental Sustainability 36, 105–115. https://doi.org/10.1016/j.cosust.2018.11.002

Wimmer, A., Kalinnik, A., Schuster, M., 2018. New insights into the formation of silver-based nanoparticles under natural and semi-natural conditions. Water Research 141, 227–234.

Windler, L., Lorenz, C., von Goetz, N., Hungerbuhler, K., Amberg, M., Heuberger, M., Nowack, B., 2012. Release of titanium dioxide from textiles during washing. Environmental science & technology 46, 8181–8188.

Wu, S., Er, M.J., Ni, M., Leithead, W.E., 2000. A fast approach for automatic generation of fuzzy rules by generalized dynamic fuzzy neural networks, in: Proceedings of the 2000 American Control Conference. ACC (IEEE Cat. No. 00CH36334). IEEE, pp. 2453–2457.

Wu, S., Zhang, S., Gong, Y., Shi, L., Zhou, B., 2020. Identification and quantification of titanium nanoparticles in surface water: a case study in Lake Taihu, China. Journal of hazardous materials 382, 121045.

Xiao, B., Zhang, Y., Wang, X., Chen, M., Sun, B., Zhang, T., Zhu, L., 2019. Occurrence and trophic transfer of nanoparticulate Ag and Ti in the natural aquatic food web of Taihu Lake, China. Environ. Sci.: Nano 6, 3431–3441. https://doi.org/10.1039/C9EN00797K

Xue, Z., Zhang, Y., Cheng, C., Ma, G., 2020. Remaining useful life prediction of lithium-ion batteries with adaptive unscented kalman filter and optimized support vector regression. Neurocomputing 376, 95–102. https://doi.org/10.1016/j.neucom.2019.09.074

Yager, R.R., 1992. Expert systems using fuzzy logic, in: An Introduction to Fuzzy Logic Applications in Intelligent Systems. Springer, pp. 27–44.

Yager, R.R., Filev, D.P., 1994. Approximate clustering via the mountain method. IEEE Transactions on Systems, Man, and Cybernetics 24, 1279–1284.

Yalezo, N., Musee, N., 2023. Meta-analysis of engineered nanoparticles dynamic aggregation in freshwater-like systems using machine learning techniques. Journal of Environmental Management 337, 117739.

Yalezo, N., Musee, N., Daramola, M.O., 2024. Developing machine learning algorithms to predict the dissolution of zinc oxide nanoparticles in aqueous environment. Environmental Nanotechnology, Monitoring & Management 101000.

Yang, X., Chen, Jishan, Shen, Y., Dong, F., Chen, Jing, 2020. Global negative effects of livestock grazing on arbuscular mycorrhizas: A meta-analysis. Science of The Total Environment 708, 134553.

Yang, X.N., Cui, F.Y., 2013. Stability of nano-sized titanium dioxide in an aqueous environment: effects of pH, dissolved organic matter and divalent cations. Water Science and Technology 68, 276–282. https://doi.org/10.2166/wst.2013.165

Yazdani, M., Zarate, P., Coulibaly, A., Zavadskas, E.K., 2017. A group decision making support system in logistics and supply chain management. Expert systems with Applications 88, 376–392.

Ye, X., Chen, B., Lee, K., Storesund, R., Zhang, B., 2020. An integrated offshore oil spill response decision making approach by human factor analysis and fuzzy preference evaluation. Environmental Pollution 262, 114294. https://doi.org/10.1016/j.envpol.2020.114294

Yeung, D.S., Tsang, E.C.C., 1997. Weighted fuzzy production rules. Fuzzy Sets and Systems 88, 299–313. https://doi.org/10.1016/S0165-0114(96)00052-8

Yildirim, Y., Bayramoglu, M., 2006. Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of Zonguldak. Chemosphere 63, 1575–1582. https://doi.org/10.1016/j.chemosphere.2005.08.070

Yin, J.-J., Liu, J., Ehrenshaft, M., Roberts, J.E., Fu, P.P., Mason, R.P., Zhao, B., 2012. Phototoxicity of nano titanium dioxides in HaCaT keratinocytes—generation of reactive oxygen species and cell damage. Toxicology and applied pharmacology 263, 81–88.

Yokel, R.A., MacPhail, R.C., 2011. Engineered nanomaterials: exposures, hazards, and risk prevention. J Occup Med Toxicol 6, 7. https://doi.org/10.1186/1745-6673-6-7

Zadeh, L.A., 2008. Is there a need for fuzzy logic? Information sciences 178, 2751–2779.

Zadeh, L.A., 1975. The concept of a linguistic variable and its application to approximate reasoning-III. Information sciences 9, 43–80.

Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8, 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zaghloul, M.S., Hamza, R.A., Iorhemen, O.T., Tay, J.H., 2020. Comparison of adaptive neuro-fuzzy inference systems (ANFIS) and support vector regression (SVR) for data-driven modelling of aerobic granular sludge reactors. Journal of Environmental Chemical Engineering 8, 103742.

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. Environment International 133, 105189. https://doi.org/10.1016/j.envint.2019.105189

Zhang, J., Prater, E.L., Lipkin, I., 2013. Feedback reviews and bidding in online auctions: An integrated hedonic regression and fuzzy logic expert system approach. Decision Support Systems 55, 894–902. https://doi.org/10.1016/j.dss.2012.12.025

Zhang, L., Li, J., Yang, K., Liu, J., Lin, D., 2016. Physicochemical transformation and algal toxicity of engineered nanoparticles in surface water samples. Environmental Pollution 211, 132–140. https://doi.org/10.1016/j.envpol.2015.12.041

Zhang, M., Yang, J., Cai, Z., Feng, Y., Wang, Y., Zhang, D., Pan, X., 2019. Detection of engineered nanoparticles in aquatic environments: current status and challenges in enrichment, separation, and analysis. Environ. Sci.: Nano 6, 709–735. https://doi.org/10.1039/C8EN01086B

Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R., 2017. Efficient kNN classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems 29, 1774–1785.

Zhang, X., Deng, Y., Chan, F.T., Adamatzky, A., Mahadevan, S., 2016. Supplier selection based on evidence theory and analytic network process. Proceedings of the institution of Mechanical Engineers, Part B: Journal of Engineering manufacture 230, 562–573.

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 in Physics 11, 625–628. https://doi.org/10.1016/j.rinp.2018.10.011

Zhang, Y., Chen, Y., Westerhoff, P., Crittenden, J., 2009. Impact of natural organic matter and divalent cations on the stability of aqueous nanoparticles. Water Research 43, 4249–4257. https://doi.org/10.1016/j.watres.2009.06.005

Zhang, Y., Chen, Y., Westerhoff, P., Hristovski, K., Crittenden, J.C., 2008. Stability of commercial metal oxide nanoparticles in water. Water Research 42, 2204–2212. https://doi.org/10.1016/j.watres.2007.11.036

Zhao, J., Lin, M., Wang, Z., Cao, X., Xing, B., 2021. Engineered nanomaterials in the environment: Are they safe? Critical Reviews in Environmental Science and Technology 51, 1443–1478. https://doi.org/10.1080/10643389.2020.1764279

Zhou, D., Keller, A.A., 2010. Role of morphology in the aggregation kinetics of ZnO nanoparticles. Water Research 44, 2948–2956. https://doi.org/10.1016/j.watres.2010.02.025

Zhou, X., Sun, Y., Huang, Z., Yang, C., Yen, G.G., 2022. Dynamic multi-objective optimization and fuzzy AHP for copper removal process of zinc hydrometallurgy. Applied Soft Computing 129, 109613. https://doi.org/10.1016/j.asoc.2022.109613

Zhu, M., Wang, H., Keller, A.A., Wang, T., Li, F., 2014a. The effect of humic acid on the aggregation of titanium dioxide nanoparticles under different pH and ionic strengths. Science of The Total Environment 487, 375–380. https://doi.org/10.1016/j.scitotenv.2014.04.036

Zhu, M., Wang, H., Keller, A.A., Wang, T., Li, F., 2014b. The effect of humic acid on the aggregation of titanium dioxide nanoparticles under different pH and ionic strengths. Science of The Total Environment 487, 375–380. https://doi.org/10.1016/j.scitotenv.2014.04.036

Zhu, Q., Wang, H., Xiao, J., 2007. Generalized dynamic fuzzy neural network-based tracking control of robot manipulators, in: International Symposium on Neural Networks. Springer, pp. 749–756.

Zhu, Y., Price, O.R., Tao, S., Jones, K.C., Sweetman, A.J., 2014. A new multimedia contaminant fate model for China: how important are environmental parameters in influencing chemical persistence and long-range transport potential? Environment international 69, 18–27.

Zimmermann, H.-J., 2011a. Fuzzy set theory—and its applications. Springer Science & Business Media.

Zimmermann, H.-J., 2011b. Fuzzy set theory—and its applications. Springer Science & Business Media.

Zimmermann, H.-J., 2010. Fuzzy set theory. Wiley Interdisciplinary Reviews: Computational Statistics 2, 317–332.

Appendices

Appendix A

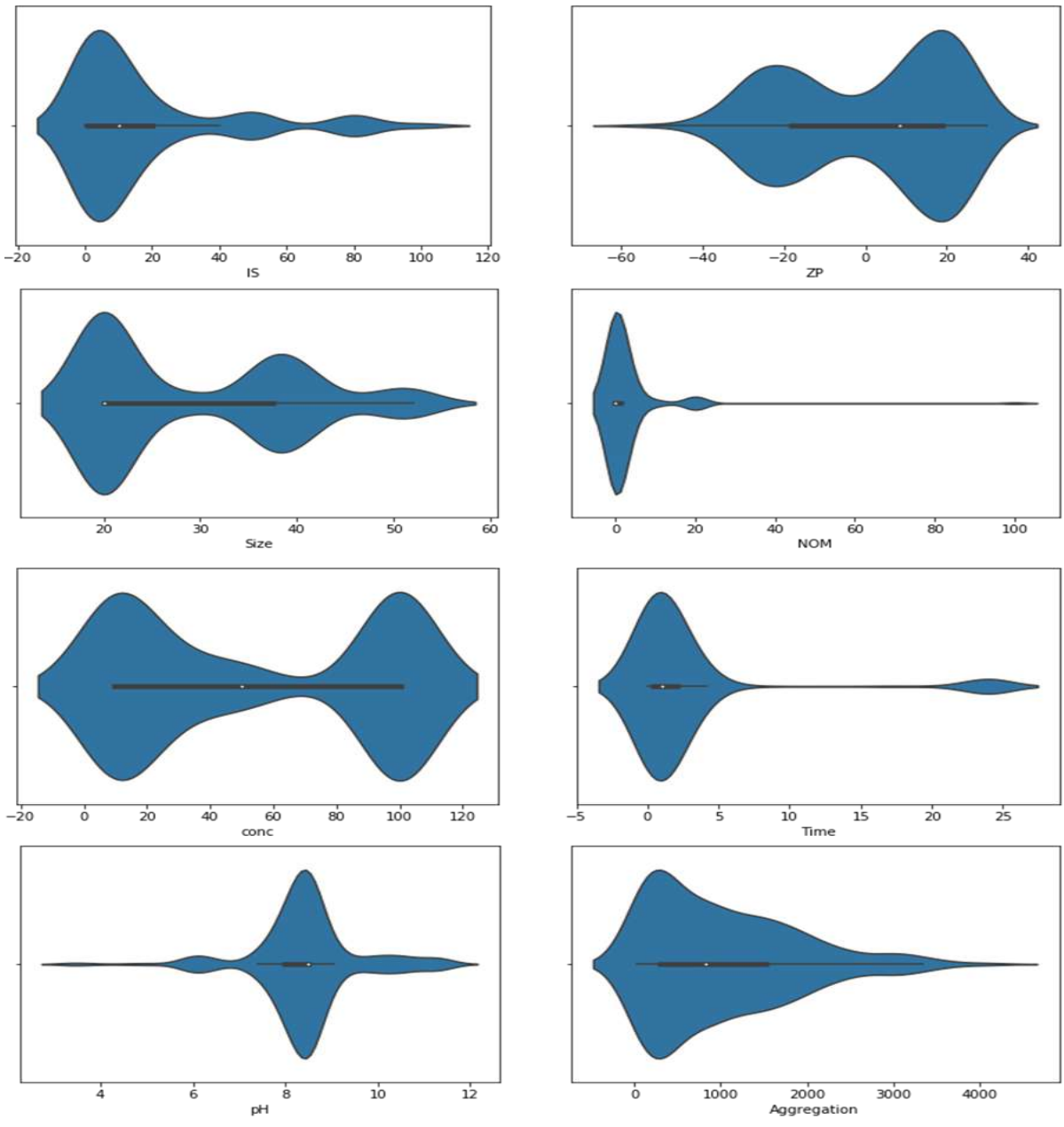


Figure A. 1 Violin plots showing the HDD data distribution in fresh-water like systems for nZnO. The white dots from violin plots, depicts the mean of each data sets, and the wide regime in these plots showed high probability distribution, whereas the skinner regime depicts low probability. The boxes bound the interquartile range (IQR) (25^{th}, 50^{th} and 75^{th} quartile). The ends of solid black points depict the highest (95^{th}) and lowest value (5th).

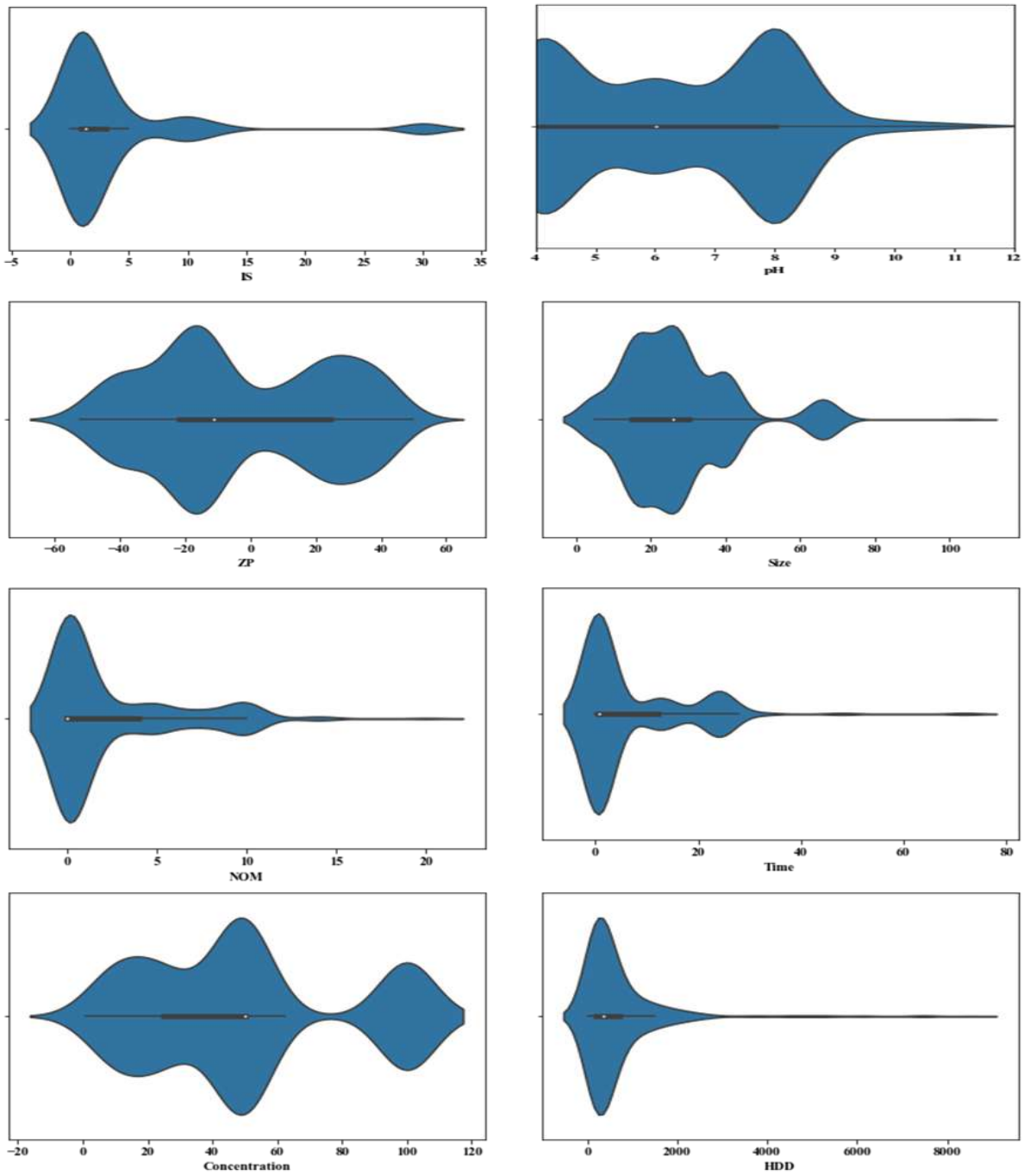


Figure A. 2. Violin plots showing the HDD data distribution in fresh-water like systems for nTiO2. The white dots from violin plots, depicts the mean of each data sets, and the wide regime in these plots showed high probability distribution, whereas the skinner regime depicts low probability. The boxes bound the interquartile range (IQR) (25th, 50th and 75th quartile). The ends of solid black points depict the highest (95th) and lowest value (5th).

Table A. 1. Literature studies on the aggregation of $nTiO_2$ and nZnO in freshwater-like systems

Type of ENPs	*Physicochemical*	*Exposure medium*	*Reference*
nTiO2	17.7 ± 6.1 nm, coated	**Purified water:** pH 5 and 7, 10 mM (KCl, $CaCl_2$), SRHA (1 mg/ℓ). $nTiO_2$ (10 mg/ℓ). Time (24 hours).	(Chowdhury et al., 2012b)
	6, 13, 23 nm, uncoated	**Purified water**: pH 7 and 10, 1, 10, and 100 mM (KCl), SRHA (1 and 10 mg/ℓ) 4. $nTiO_2$ (1 mg/ℓ). Time (0, 0.139, 0.27, 0.417, 1.11 hour).	(Chowdhury et al., 2013)
	27 ± 4 nm, 51.5 ± 3 m^2/g, uncoated	**Mesocosm freshwater**: pH 8.38, 7.18 eq. ℓ, TOC (5.283 µM). nTiO2 (10, 50, 100 mg/ℓ). Time (0-1 hour).	(Keller et al., 2010)
	5 nm, anatase, uncoated	Purified water: pH 8, Ca = 3 and 0.25 mM, SRHA (3.81 and 8.81 mg/ℓ) 4. $nTiO_2$ (10, 20, 30, 40, 50 and 100 mg/ℓ). River water; pH 7.8, DOC, (2 mg/ℓ), Ca (0.244 mM), NO_3 (0.055 mM), SO_4 (0.103 mM), Cl (0.356 mM), Na (0.415 mM), nTiO2 (10 mg/ℓ). Time (0.5 hours)	(Adam et al., 2016)
	10nm x 40 nm, uncoated, 130-190 m^2/g	**Purified water;** pH 4, 6, and 8, HA: 0–210 µg/ℓ., TiO_2 (20 mg/ℓ), NaCl (0 to 4.5 mol. ℓ). Time (1 hour)	(M. Zhu et al., 2014b)
	30 nm, 51.5 ± 3 m^2/g uncoated	**Purified water**: pH 8, $CaCl_2$ (0.01, 0.05,1, 10, mM), SRHA (10 mg/ℓ). Time (0-0.5 hour). $nTiO_2$ (100 mg/ℓ).	(Thio et al., 2011)
	15 nm, 17 mV uncoated	**Tap water**: pH =8.1 ± 0.2, conductivity = 750-940 uS/cm, IS = 0.01 M, TOC (0-8.9 mg/ℓ). Time (0 – 8 hours). $nTiO_2$ (10 mg/ℓ).	(Zhang et al., 2008)
	15 nm, 240 m^2/g, uncoated anatase	**Purified water**: pH 4.5, 6.2, and 9.8- 10, IS = 0.001 M (NaCl), SRHA (1- 10 mg/ℓ). Time (0-70 hours). $nTiO_2$ (50 mg/ℓ).	(Loosli et al., 2013)
	25 nm, uncoated pure anatase. 16.2 mV	**Lake water**: pH 6.5, Cl (22.59), NO3 (1.17), SO4(16.62), Ca (47.3), Mg (2.92) mg/ℓ, TOC (14.0 mg/ℓ). Time (0- 10 hours), $nTiO_2$ (1 mg/ℓ).	(Tso et al., 2010)

19.8 nm, 57 m^2/g uncoated	**Given order below**: Water type, pH, IS (mM) and DOC (mg/ℓ), and Zeta potential (mV) **EPA very hard water:** 8.4, 9.6, < 0.5, and -10 ± 0.2; **EPA moderately hard water**: 7.4, 9.6, < 0.5 and -16 ± 0.6; **EPA very soft water**: 6.4, 0.3, < 0.5 and -20 ± 1.6; **Tap water:** 8.0, 1.8, 1.03 and -9.0 ± 0.3; **Lake water:** 7.9, 1.6, 2.12 and -15 ± 0.6; **Groundwater:** 7.8, 3.5, 1.57 and -1 ± 0.6; **Peat bog water:** 5.2, 0.5, 37.2, and -26 ± 0.2. nTiO2 (25 mg/ℓ). Time (15 hours).	(Ottofuelling et al., 2011)
13.3± 1.3 nm, 206 m^2/g, uncoated	**Purified water**: pH 2.8, (10 mM NaCl), SRFA (16-100uM uM). Short-term SRFA (0.9–36.4 µM), long-term SRFA (0.9–10.9 mM), and Time (0 -168 hours of investigation. $nTiO_2$ (104 mg/ℓ).	(Danielsson et al., 2017)
5-30 nm, uncoated, rutile	**Purified water**: pH 7.0 ± 0.2, IS (1 mM (0,1, 2, 4, 6, 8 Ca mg/ℓ)), DOM (0,1, 2, 4, 6 and 10 mg/ℓ). Time (24 hours). $nTiO_2$ (10 mg/ℓ).	(Yang and Cui, 2013)
15 nm, uncoated	**Purified water**: pH 7.8 ± 0.2, 0.01 M KCl NOM (0-10 mg/ℓ). Time (2 hours). $nTiO_2$ (10 mg/ℓ).	(Zhang et al., 2009)
P25: 30 nm, 35-45 m^2/g, uncoated	**Purified water**: pH 4, 6, and 8, 5 mM (NaCl), FA (0-5 mg/ℓ). Time (24 hours). $nTiO_2$ (50 mg/ℓ).	(Li and Sun, 2011)
21 nm, uncoated	**Given order below**: water type, pH, IS (mM), DOC (mg C/ℓ), Ca^{2+} (mM) **Lake water:** 7.08, 3.46, 14.66 and 0.41. **Groundwater:** 6.53, 14.90, 5.12 and 1.48. **River water:** 6.89, 1.96, 15.76 and 0.06. nTiO2 (25 mg/ℓ). Time (1 hour).	(Chekli et al., 2015)
21 nm, mixture of anatase and rutile uncoated,	**Purified water**: pH 3 – 9, NaCl 0.09 mM, $CaCl_2$ 0.3, $MgCl_2$ 0.3 mM, HA (1 mg/ℓ). Time (1.17 h). $nTiO_2$ (100 mg/ℓ). IS= 0.9 mM	(Romanello and Fidalgo de Cortalezzi, 2013)
13. 5 ±1.2 nm, 162 m^2/g,	**Purified water**: pH 5, I= 10 mM NaCl), SRFA (0-160 uM) and (0-100uM) Time (0 -168 hours). $nTiO_2$ (25 mg/ℓ).	(Danielsson et al., 2018)
15 nm, 240 m^2/g, uncoated anatase	**Purified water**: pH 2 - 10, I= 0.001 M, SRFA (3-10 mg/ℓ). Time (0-70 hours). $nTiO_2$ (50 mg/ℓ).	(Loosli et al., 2014)
5 nm, 5 248 ± 4 m^2/g, uncoated. 25 nm,	**Purified water**: pH 3.7, 6,0 and 8, 0, I (0.02 M NaCl), SRHA (0, 10, 20, and 50 mg/ℓ). Time (24 hours). $nTiO_2$ (100 mg/ℓ).	(Jayalath et al., 2018)

	15 nm, uncoated,	**Purified water**: pH 2 – 10, 6, 8, I = 0.001 M (NaCl), FA (0, 2, 4,8,12 mg/ℓ). Time (0 – 2.33 hours). $nTiO_2$ (50 mg/ℓ).	(Palomino et al., 2013)
	25.1 ±8.2 nm, 51.1 m^2/g, uncoated.	**Purified water**: pH 7.8 ± 0.1, I= 3, 3.4, 6.8, 10, and 13.6 mM, SRNOM (0 and 5 mg/ℓ). Time (0, 25, 0.5, 1, 2, 4, 6, 8, 24, 28, 48, 76, 96 hours), $nTiO_2$ (0.1 to 5 mg/ℓ).	(L. Li et al., 2016)
	52 ±19 nm, anatase	**Given in the following order:** water type, pH 3-4, IS (0-500 (meq/ℓ) NaCl), SRHA (1, 10, 15, 20, 25, 50 mg C/ℓ), Ca^{2+} (mM) $nTiO^2$ (100 mg/ℓ). Time (0-2 hours).	(Hsiung et al., 2016)
	20-30 nm, 46 m^2/g, uncoated	**Purified water**: pH 6, IS (CaCl =10 and NaCl = 10 mM), HA (1 - 10 mg/ℓ). Time (24 hours). $nTiO_2$ (50 mg/ℓ).	(Liu et al., 2013)
	40nm, 50±15 m^2/g, uncoated.	**Given order below**: water type, pH, DOC (mg C/ℓ), IS; NOx, NH_4, P(mg/ℓ) Brackish water (BW): **8, 9.1,** 0.004,0.011,0.022 Nutrient-rich lake (NR): 7.7, 8.2, 0.052,0.035, 0.0082, Humus-poor lake (HP): 6.7, 8.7, 0.237, 0.036,0.022 Humic lake (HL): 6.7, 8.1, 0.001, 0.003, 0.066 nTiO2 (10 mg/ℓ). Time (3 hours).	(Li et al., 2016)
	20 ±5 nm, rutile	**In the order below**: water type, pH 1-10, IS (CaCl2 (10.7 mM), NaCl (1380 mM)) SRFA (10 and 100 mg C/ℓ), $nTiO_2$ (20 mg/ℓ). Time (0, and 1 hours).	(Raza et al., 2016)
	21 nm, surface area 61 m^2/g, uncoated.	**Given order below**: water type, pH, IS (mM), DOC (mg C/ℓ), **Freshwater:** 8.14, 2.0, and < 0.1. **Estuarine water:** 8.06, 9.8, and < 0.1 **Lagoon water:** 8.17, 984, and 2.16. nTiO2 (0.1, 1.0, and 10 mg/ℓ). Time (0.2, 25, and 50 hours).	(Brunelli et al., 2013)
nZnO	20.7 ± 6.1 nm, uncoated 50±15 m^2/g,	**Purified water2:** pH 8.4, 9.3. and 10.4, IS 10-3 M NaCl), SRHA (0-0.5 mg/ℓ). nZnO (100 mg/ℓ). Time (1 hour).	(Mohd Omar et al., 2014)
	18.4 ± 6.0 nm, uncoated	**Purified water**: pH 5 and 7, 1 and 10 mM (KCl), 25°C. NOM (0.0135 mg/ℓ) 4. nZnO (100, 200, 400, and 800 mg/ℓ). Time (1.66 hours).	(Han et al., 2014)
	24 ± 3 nm, 42.1 ± 3 m^2/g, 100% zincite hexagonal, uncoated	**Mesocosm freshwater**: pH 8.38, 7.18 eq. ℓ, TOC (5.283.45µM). nZnO (10 mg/ℓ). Time (0 – 1 hour).	(Keller et al., 2010)

50-70 nm, uncoated	**Purified water**: pH 7.8 ± 0.2, 0.01 M KCl NOM (0-10 mg/ℓ). Time (2 hours). nZnO (10 mg/ℓ).	(Zhang et al., 2009)
20 nm uncoated,	**Lake water**: pH 8-11, I (0-120 mM NaCl), NOM (0-100 mg/ℓ). Tim (0-16.67 hours). nZnO (100 mg/ℓ).	(Zhou and Keller, 2010)
37.5 nm, uncoated	**Purified water**: pH 7.82, 8.77, 9.90, 10.56, and 11.26, IS (1 to 200 meq/ℓ), (NaCl and $NaSO_4$), SRHA (0-100 mg/ℓ). Time (0-16.67 hours). nZnO (100 mg/ℓ).	(Y.-H. Peng et al., 2017c)
50-60, uncoated	**Tap water**: pH =8.1±0.2, IS = 0.01 M, TOC (0-8.9 mg/ℓ). Conductivity (750 -940 uS/cm1 Time (0 – 8 hours). nZnO (10 mg/ℓ).	(Zhang et al., 2008)
20 nm, -23 mV, uncoated	**Lake water**: pH 6.5, Cl (22.59), NO3 (1.17), SO4(16.62), Ca (47.3), Mg (2.92) mg/ℓ, $22.5°C$. TOC (14.0 mg/ℓ). Time (1-10 hours) contact time. nZnO (1 mg/ℓ).	(Tso et al., 2010)
15 nm, 240 m^2/g, uncoated	**Purified water**: pH 2- 10, IS = 0.001 M (NaCl), HA (3, 4, 5, 7 and 10 mg/ℓ). Time (0-70 hours, nZnO (50 mg/ℓ).	(Jiang et al., 2012)
28 +-11 and 21 -+8 nm	**Purified water:** pH 8.5-8.6, I= 0.1 M (80 mM KCl), SRHA, and SRFA (0, 1, 5, 10, 20 mg/ℓ). Time (0-80 min) contact time. nZnO (75 uM).	((Jiang et al., 2015b)
52 nm, 19± 3 m^2/g, uncoated	**Given order below**: water type, pH, IS (mM), TOC (mg /ℓ), very soft water**:** 6.4-6.8, 0.572, 2 moderately hard water; 7.4-7.8, 4.58, 2 very hard water; 8.0-8.4,18.34, 2 nZnO (20 mg/ℓ). Time (0 - 180 hours)	(Majedi et al., 2014b)
41± 8 nm, 26± 3 m^2/g,	**Purified water**: pH 6, IS (2 mM KCl), HA (5 mg/ℓ). Time (0-3 hours). nZnO (100 mg/ℓ).	(Majedi et al., 2014c)
20±5 nm (99.5), uncoated, 50 m^2/g	**The following order is given:** water type, pH, IS (mM), TOC (mg /ℓ), Pond(TX4)**:** 7.06, 6.88, 15.4 countryside river(DY2); 7.30, 3.19, 23.3 Dongjiang river (SX1) 7.19, 5.19, 7.74 Taihu lake (TH3); 7.17, 7.99, 15.1 nZnO (100 mg/ℓ). Time (0.5 hours).	(L. Zhang et al., 2016)

<100nm. uncoated	**Natural water**: pH 7.7 +- 0.1, DOC (4.65 +- 0.23 mg/ℓ), Conductivity (242 _+9.0 6 uS/cm^{-1}) Time (24 hours). nZnO (10 and 25 mg/ℓ).	(Fang et al., 2017)
20 nm, 30-50 m^2/g, uncoated	**Purified;** pH 7- 14, HA and FA (1 mg/l), Time (6 hours). C(10 mg/ℓ).	(Jones and Su, 2014)
10-15 nm, uncoated	**Given in the following order:** water type, pH, IS (mM), TOC (mg C/ℓ), IS (Cl, SO^{4-}, Ca^{2+}, Mg^{2+}, Na^{+} mg/ℓ), lake water**:** 8.3, 5.1, 3.4, 22, 44, 4.6, 2.7 Tap water; 8.2, 10, 11, 29, 66, 7.8, 6.7 Time (0, 48, and 144 hours). nZnO (10 mg/ℓ).	(Muna et al., 2018)
<50 nm, crystalline structure, Hexagonal, 12.5 m^2/g, uncoated	**Freshwater**: pH 6.90, IS 0.79 mM, TOC (10 mg/ℓ), conductivity (119 us cm^{-1}. Time (0-0.55 hours). nZnO (10 mg/ℓ).	(Khan et al., 2019)
40 to 50 nm 19.8 nm, 57 m^2/g, uncoated	**Given in the following order:** water type, pH, IS (mM), TOC (mg C/ℓ), Lake water**:** 7.31±0.43, 6.11±0.42, 11.28±0.93 Tap water; 7.52±0.37, 4.15±0.36, 6.65±0.48 Rainwater; 7.66±0.32, 3.66 ±0.23 21.26±1.83 Pool water; 7.96±0.24, 6.96±0.52, 15.64±1.40 nZnO (100 mg/ℓ). Time (2, 20, 40, 60, 80, 100, 2 hours).	(Liu et al., 2018)

*The freshwater used (TOC; Total dissolved carbon, and DOC; dissolved organic carbon) and synthetic medium were based on (SRHA; Suwannee River Humic acid, and SRFA; Suwannee River Fulvic Acids) as surrogates for natural organic matter (NOM) as recommended by the OECD test guideline to estimate the stability of ENPs (Monikh et al., 2018).

Table A. 2. Descriptive statistics for normalised data sets of $nTiO_2$ and nZnO

		nZnO		$nTiO_2$	
Variables	**Units**	**Mean**	**SD**	**Mean**	**SD**
NOM	$\mathrm{mg} \cdot \ell^{-1}$	0.03	0.09	0.12	0.18
pH	-	0.62	0.14	0.42	0.22
IS	mM	0.17	0.24	0.11	0.20
Size	nm	0.29	0.34	0.29	0.20
ZP	mV	0.25	0.65	0.49	0.27
ENP Concentration	$\mathrm{mg} \cdot \ell^{-1}$	0.49	0.46	0.48	0.31
Time	hour	0.11	0.24	0.10	0.16
HDD	nm	0.24	0.21	0.08	0.12

Table A. 3. Performance parameters of the prediction models on the aggregation of $nTiO_2$ for the calibration and validation sets.

				RMSE		R	
Models		Combination		Train	Test	Train	Test
ANFIS	1	GP	Triangular	0.05	0.10	0.93	0.79
	2		Trapezoidal	0.07	0.12	0.86	0.67
	3		Generalized Bell	0.04	0.15	0.95	0.59
	4		Gaussian-I	0.03	0.15	0.97	0.57
	5		Gaussian-II	0.04	0.10	0.95	0.76
ANN	1	Adam	Tanh	0.06	0.09	0.84	0.83
	2		Sigmoid	0.07	0.10	0.83	0.82
	3		ReLU	0.10	0.12	0.64	0.63
	4	SGD	Tanh	0.10	0.13	0.60	0.58
	5		Sigmoid	0.11	0.14	0.50	0.42
	6		ReLU	0.11	0.13	0.40	0.39
RFR	1	DT,	(20^a, 42^b)	0.03	0.07	0.98	0.93
	2		(60^a, 42^b)	0.02	0.08	0.99	0.91
	3		(100^a, 42^b)	0.02	0.08	0.99	0.92
SVM	1	RBF	(1^c, 0.1^d,1^e)	0.10	0.12	0.63	0.65
	2		(1^c, 0.3^d,1^e)	0.22	0.19	0.60	0.62
	3		(1^c,0.1^d,10^e)	0.07	0.09	0.87	0.85
	4	Poly	(1^c, 0.1^a,1^b)	0.11	0.13	0.51	0.47
	5		(1^c, 0.3^d,1^e)	0.18	0.17	0.48	0.46
	6		(1^c,0.1^d, 10^e)	0.11	0.13	0.54	0.51
MLR		-	---	0.13	0.15	0.27	0.14

DTs: decision trees, SDG: stochastic gradient descent, Adam: adaptive momentum, GP: Grid Partition, ReLu: Rectified linear unit; DTs: a, Randomized state: b C (cost of constraint violation) : c, ε (epsilon): d, γ (gamma):e, RBF: radial basis function, Poly: polynomial.

Table A. 4. Performance parameters of the prediction models on the aggregation of nZnO for the calibration and validation sets.

				RMSE		R	
Model		Combination		Train	Test	Train	Test
ANFIS	1	GP	Triangular	0.07	0.28	0.94	0.54
	2		Trapezoidal	0.06	0.24	0.96	0.60
	3		Generalized bell	0.06	0.28	0.96	0.50
	4		Gaussian-I	0.06	0.29	0.96	0.49
	5		Gaussian-II	0.06	0.20	0.96	0.70
ANN	1	Adam	Tanh	0.09	0.11	0.93	0.90
	2		Sigmoid	0.08	0.12	0.93	0.90
	3		ReLU	0.11	0.15	0.77	0.75
	4	SGD	Tanh	0.17	0.20	0.72	0.65
	5		Sigmoid	0.14	0.18	0.55	0.54
	6		ReLU	0.19	0.21	0.52	0.50
RFR	1	DTs	20^{a}, 42^{b}	0.03	0.11	0.99	0.91
	2		60^{a}, 42^{b}	0.03	0.11	0.99	0.91
	3		100^{a}, 42^{b}	0.03	0.11	0.99	0.91
SVM	1	RBF	(1^{c}, 0.1^{d},1^{e})	0.13	0.16	0.81	0.75
	2		(1^{c}, 0.3^{d},1^{e})	0.19	0.20	0.71	0.67
	3		(1^{c},0.1^{d},10^{e})	0.08	0.12	0.93	0.89
	4	Poly	(1^{c}, 0.1^{d},1^{e})	0.20	0.23	0.39	0.36
	5		(1^{c}, 0.1^{d},3^{e})	0.20	0.22	0.41	0.40
	6		(1^{c},0.3^{d},10^{e})	0.20	0.24	0.40	0.29
MLR		-	---	0.21	0.23	0.33	0.33

DTs: decision trees, SDG: stochastic gradient descent, Adam: adaptive momentum, GP: Grid Partition, ReLu: Rectified linear unit; DTs: a, Randomized state: b C: c, ε: d, γ :e, RBF: radial basis function, Poly: polynomial.

Appendix B

Table B. 1. Literature studies on the dissolution of nZnO in freshwater-like systems

Physicochemical	*Exposure medium*	*Reference*
1. 6 nm and spherical 2. 20 nm and spherical 3.71 nm irregular Hexagonal wurtzite, uncoated	**Purified water:** pH 7−9, Conc. (50−100 mg/ℓ). Time (0-200 mins). Ionic strength = 0.1 M, buffer concentration of 0.02M (KCl or KNO_3)	David et al., 2012
5 nm zinc oxide, coating type: polyvinylpyrrolidone (PVP)	**Purified water**: pH 4 and 7.4, Conc. (0.01366 mg/ℓ), Time (24 hours).	Briffa, 2018
PEG; PVA- and PVP- and bare ZnO were 58 ± 6.3, 60 ± 1.9, 52 ± 2.3 and 69 ± 1.2 nm respectively. platelets	**Purified water:** pH 7, 1 mM NaHCO3. SRHA (0, 5, 20 and 80 mg/ ℓ) Conc. (20 mg/ℓ), time (1 to 72 hours)	Kizhakkumpat et al., 2021
28±11 nm, 38m^2/g, cationic 3-aminopropyl triethoxysilane, wurtzite-like crystal structure	**Purified water**: pH 8.5-8.6. 0.1 M KCl SRHA, SRFA (1 to 40 mg/ℓ). Time (0-1 hour). Conc. (75 Mm/20 mM). 25 ºC	Jiang et al., 2015
20 ± 5 or , 19 ± 7, 15.0 ± 0.87 mv, 50 ± 10 m^2/ g	**Given order below**: water type, pH, IS (mM), DOC (mg C/ℓ), Conductivity mS/cm Davis medium: 7.2, 0.04, 495, 3.47 Luria-Bertani medium: 7.0, 0.198, 6216, 21.8. Conc. (50 and 100) mg/ℓ, time (50 hours)	Li et al., 2011
20 nm, uncoated	**Purified water**: pH 6.0, 7.0, 8.0, 9.0, TOC for SDS and NP-9 stock solutions were 537.3 ± 4.2 and 616.8 ± 5.8 mg/ℓ, respectively. Time (0-24 hours). Conc. (50 mg/ℓ).	Li et al., 2017
7, 17, 24, 15 nm 47 and 130 nm, wurtzite crystal structure and spherical in shape. Not coated	**Purified water**: pH = 7.5, 0 – 500 mM) (NaCl), Citric acid (0, 5, 10, 50 , 100 mM). Time (24 hours), $nTiO_2$ (500 mg/ℓ).	Rupisana et al., 2011
7, 17, 24, 15 nm 47 and 130 nm, wurtzite crystal structure for the ZnO. Spherical in shape and not uncoated	**Purified water**: pH 7.5, Citric acid (0, 5, 10, 25, 50 mM). Time (24 hours), $nTiO_2$ (500 mg/ℓ).	Mudunkothuwa et al., 2012

4.5 and 27 nm, coated, Citrate, Gelatin, PVP and Chitosan	**Purified water**: pH 7.6 and 6.1, IS: 0.029 M. (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, NH_4^{+}, $NO3^{-}$, $SO4^{2-}$, $HCO3$), DOM (0, 1, 2, 4, 6 and 10 mg/ℓ). Time (5, 10, and 19 days). Conc. (1.0 mg/ℓ) in dark at 23^{o}C	Odzak et al., 2014
4, 15 and 241 nm, 105±13 m^2/g, wurtzite structure, uncoated	**Purified water**: pH 1.0, 3.0 and 6.0, 9.0, and 11. IS= 0, 0.02 M (NaCl), Humic acid (100 mg/ℓ). Time (24 hours). Conc. (0.100 and 1000 mg/ℓ).	Bian et al., 2011
20 ± 5a, 7, 15.0 ± 0.87 mv, 50 ± 10, m^2/ g	**Given order below**: water type, pH, IS (mM), DOC (mg C/ℓ), **Tap water:** 7.5, 4.0, 6.8 **West lake:** 7.8, 4.3, 12.6 **Xixi river:** 7.5, 5.7, 5.3 **Qiantang river:** 7.9, 7.5, 6.1 Conc. (0-50 mg/ℓ).	Li M et al., 2013
Uncoated ZnO-NPs (20 nm), dodecyltrichlorosilane-coated ZnO-NPs (23 ± 1 nm) and 3-aminopropyltrimethoxysilane ZnO-NPs (24 ± 1 nm)	**Purified water**: pH 7 or 8 5, IS= 0.1 mM (NaCl), Time (0-96 hours). Conc. (0.1, 0.5, 1, 3, 5, 10, 30, 50, 80 or 100 mg/ℓ). 25 ± 1 °C	Yung et al., 2017
Polyacrylic acid-stabilized (PAA) nZnO-(20 nm), Sodiumhexametaphosphate-stabilized, (HMP) (40 nm)	**Purified water**: pH 7.0, I= 0.01M, Time (24 h), Conc. (5.0 mg/ℓ).	Merdzan et al., 2014
50 nm and < 100 nm hexagonal wurtzite structure. Not coated.	Purified water: pH =1, 6, 7 and 9), I= (3 mM and 5 mM) (NaCl). Time (24 hours). Conc. (100 mg/ℓ).	Velintine et al., 2017
18.3 nm, not coated.	**Purified water**: pH 8.5, IS = 5.6 mM, TOC = 219 mg/ℓ, conductivity = 318 µs/cm. Time (0-72 hours), Conc. (0.1 and 1 mg/ℓ).	Li et al., 2011
35 nm, not coated.	**Given order below**: pH, TOC (mg C/ℓ), IS (mg/ℓ), Conductivity (µS/cm) Laboratory: **8.0,** 2.1**,** <1, 310 Brown: 7.1, **6.5, 5, 82.5** Green: 8.1, 2.9, 13, 277 Time (1, 48, and 96 hours). Conc. (10 mg/Lmg/ℓ).	Gagné et al ., 2019

70 nm, not coated.	**Given order below**: water type, pH, DOC (mg C/ℓ), IS; Ca^{2+}, Cl^{-}, SO_4^{2-} (mg/ℓ) River 1: 8.2, 13.3, 122 15.4, 96.1 River2: 8.2, 13.2, 124, 17, 69.1 River 3: 7.9, 25.9, 106, 4.6, 55.6 River 4: 8.1, 29.2, 111, 13.5, 76.1 River 5: 7.5, 34.5, 82.0, 9.2, 20.9 River 6: 8.1, 31.5, 58.0, 7.7, 14.9 Conc. (0.001, 0.01, 0.1, 1, 10 mg/ℓ). Time (24 hours).	Blinova et al., 2010
14.9 ± 4.5 nm, not coated	**In the order below**: pH 8.12, IS= 1.02 mg /ℓ (Nitrate), DOC (0.65 mM), Conc. (10, 100, 1000 mg/ℓ). Time (0, and 1 hours).	Du et al., 2019
<_130 nm, not coated	Given order below: water type, pH, IS (mM), DOC (mg C/ℓ), Conductivity mS/cm Lake Greifen Maur: 8.1, 6.4, 3.1, and 404 River Rhine Eglisau: 8.2, 5.3, 1.9 and 318 Lake Lucerne Kastanienbaum: 8.2, 3.4, 1.1, and 202 Lake Gruere Saignelegier: 6.5, 1.4, 21.7 and 96 Lake Cristallina Alps Ticino: 6.4, <0.3, 1.0, and 7.0 Rainwater Dübendorf: 4.8, <0.3, 0.9 and 15.5 Time (0, 1, 3, and 8 days). 22.5 ± 0.1 ^{o}C) Conc. (50 mM).	Odzak et a., 2017
Not coated crystallized with wurzite structure. Shape of nZnO-(137.9 nm) was irregular, and nZnO-2 (19.2 nm) was spherical.	**Purified water**: pH (6.0 and 9.0), SRHA (1 and 5 mg/ℓ). Conc. (20 mg/ℓ), Time (0.5 hours).	(Han et al., 2014)
20 nm, spherical, not coated	**Purified water**: pH (4–10), IS (0.005–0.1 M), SRFA (0–60 mg/ℓ), Time (24 hours). Conc. (1.0–30.0 mg/ℓ).	Domingos et al., 2013
30-30 nm, not coated, spherical, 25.7 ± 0.3 mV	**River water**: pH = 7.9, IS = 0.008 M, DOC (35.37 mg/ℓ) Time (24, 48, and 72 hours). Conc. (70 ug/ ℓ).	Londono et al., 2017
15 nm, 240 m^2/g, not coated	**Purified water**: pH =2-10, IS = 0.001 M (NaCl), HA (3, 4, 5, 7 and 10 mg/ℓ). Time (0-70 hours), Conc. (50 mg/ℓ).	(Jiang et al., 2012)

52 nm, 19± 3 m^2/g, not coated	**Given order below**: water type, pH, IS (mM), TOC (mg /ℓ), very soft water**:** 6.4-6.8, 0.572, 2 moderately hard water; 7.4-7.8, 4.58, 2 very hard water; 8.0-8.4,18.34, 2 Conc. (20 mg/ℓ). Time (0 - 180 hours)	(Majedi et al., 2014b)
41± 8 nm, 26± 3 m^2/g,	**Purified water**: pH 6, 7.5, and 9.0, IS (2 mM KCl), 50 mM $CaCl_2$, 10 Mm NaNO3), HA (5 mg/ℓ), Oxalic acid (2 mg/ℓ), Citric acid (20 mg/ ℓ). Time (0-50 hours). Conc. (100 mg/ℓ).	(Majedi et al., 2014c)
20±5 nm (99.5), not coated, 50 m^2/g	**The following order is given:** water type, pH, IS (mM), TOC (mg /ℓ), Pond**:** 7.06, 6.88, 15.4 countryside river ; 7.30, 3.19, 23.3 Dongjiang river: 7.19, 5.19, 7.74 Taihu lake; 7.17, 7.99, 15.1 Conc. (100 mg/ℓ). Time (0.5 hours).	(L. Zhang et al., 2016)
10-15 nm, not coated	**Given in the following order:** water type, pH, IS (mM), TOC (mg C/ℓ), IS (Cl, SO^{4-}, Ca^{2+}, Mg^{2+}, Na^{+} mg/ℓ), lake water**:** 8.3, 5.1, 3.4, 22, 44, 4.6, 2.7 Tap water; 8.2, 10, 11, 29, 66, 7.8, 6.7 Time (0, 48, and 144 hours). Conc. (10 mg/ℓ).	(Muna et al., 2018)
<50 nm, crystalline structure, Hexagonal, 12.5 m^2/g, not coated	**Freshwater**: pH=6.95, IS = 0.002 mM, conductivity (82.4 us cm^{-1}). Time (0.55 -10 days). Conc. (10 mg/ℓ).	(Khan et al., 2019)
40 to 50 nm, 57 m^2/g, not coated	**Given in the following order:** water type, pH, IS (mM) TOC (mg C/ℓ), Lake water**:** 7.31 ± 0.43, 6.11 ± 0.42, 11.28 ±0.93 Tap water; 7.52 ± 0.37, 4.15 ± 0.36, 6.65±0.48 Rainwater; 7.66 ± 0.32, 3.66 ± 0.23 21.26 ±1.83 Pool water; 7.96 ±0.24, 6.96 ±0.52, 15.64 ±1.40 Conc. (100 mg/ℓ). Time (2, 20, 40, 60, 80, 100 hours).	(Liu et al., 2018)

*The freshwater used (TOC; Total dissolved carbon, and DOC; dissolved organic carbon) and synthetic medium were based on (SRHA; Suwannee River Humic acid, and SRFA; Suwannee River Fulvic Acids, Oxalic acid, Citric acid) as surrogates for natural organic matter (NOM)

Table B. 2. Formalisation of categorical inputs for application of one hot coding

Inputs	Categories
NOM types	1. DOC
	2. SRHA
	3. SRFA
	4. TOC
Salt types	1. Monovalent (Na^+, Cl^-, HCO_3^- K , NH_4+, NO_3^-)
	2. Divalent (Ca^{2+}, SO_4^{2-}, Mg^{2+}, SO_4^{2-})
	3. Both (Na^+, Cl^-, HCO_3^- Ca^{2+}, SO_4^{2-}, Mg^{2+}, SO_4^{2-}, K^+)
Coating	1. Coating (dodecyltrichlorosilane, and 3-aminopropyltrimethoxysilane, Citrate, Gelatin, PVP and Chitosan, PAA, HMP, cationic 3-aminopropyl triethoxysilane, PEG, PVA- PVP)
	2. Non-coating
Coating types	1. Bare
	2. Electrostatic e.g. Citrate,
	3. Steric e.g. PEG, PVA Polyvinyl alcohol l(PVA), PVP
Shapes	1. Spherical
	2. Non-spherical

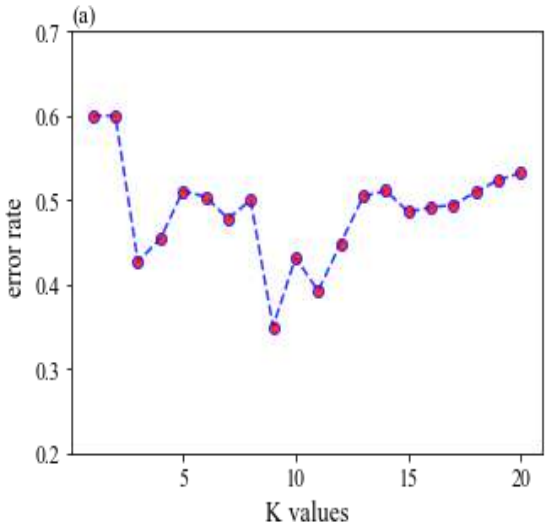


Figure B.1. k values in KNN

Appendix C

Table C. 1. Sub-criteria to score intermediate parameters

Transformation	***Qualitative values***	**weights**
Aggregation	*very high*	9
	high	*7*
	medium	*5*
	low	*3*
	very low	*2*
	extreme low/none	1
Dissolution	*extreme low/none*	1
	very low	2
	low	3
	medium	4
	high	5
	very high	6
Diss *Final*	*extreme low/none*	1
	very low	2
	low	3
	medium	5
	high	7
	very high	9
Stabilisation	*extreme low/none*	1
	very low	2
	low	3
	medium	5
	High	7
	very high	9

Appendix D

Table D. 1. Samples of linguistic rules to estimate different outputs

Rules	IF		AND		AND		THEN	
	Input	MF	Input	MF	Input	MF	Output	MF (weights)
Linguistic rules for estimating the WC-driven aggregation (WCA)								
R_1	NOM	*low*	pH	*high*	IS	*low*	WCA	*low (1)*
R_2	NOM	*low*	pH	*low*	IS	*low*	WCA	*high(1)*
R_{27}	NOM	*moderate*	pH	*low*	IS	*low*	WCA	*moderate (0.5)*
Linguistic rules for PC-driven aggregation (PCA)								
R_1	Coating	*low*	ZP	*low*	*Size*	*low*	PCA	*high(1)*
R_2	Coating	*moderate*	ZP	*low*	Size	*low*	PCA	*moderate (1)*
R_{27}	Coating	*high*	ZP	*low*	Size	*low*	PCA	*low (0.50)*
Linguistic rules for effective aggregation (EA)								
R_1	WCA	*low*	PCA	*moderate*	-	-	EA	*low (1)*
R_7	WCA	*very high*	PCA	*high*	-	-	EA	*low (0.75)*
R_{25}	WCA	*very high*	PCA	*very high*	-	-	EA	*very high(1)*
Linguistic rules for collision efficiency (CE)								
R_1	EA	*low*	EC	*high*	-	-	CE	*low (1)*
R_7	EA	*high*	EC	*high*	-	-	CE	*very high (1)*
R_{25}	EA	*very high*	EC	*low*	-	-	CE	*low (1)*
Linguistic rules for dispersion using ionic quantity (IQ) **and particle quantity** (PQ)								
R_1	IQ	*high*	PQ	*low*	-	-	Dispersion	*very low (1)*
R_7	IQ	*low*	PQ	*high*	-	-	Dispersion	*high (1)*
R_{36}	IQ	*very high*	PQ	*very low*	-	-	Dispersion	*extremely low(1)*

Table D. 2. Membership function and respective values of intermediate input variables

Intermediate	***Range***	***Mfs***	***Fuzzy numbers***
EA, ES, ED	[0 1]	Very Low	(0, 0, 0.1, 0.2)
		Low	(0.1, 0.3, 0.4)
		Moderate	(0.35, 0.5, 0.6)
		High	(0.55, 0.7, 0.9)
		Very high	(0,8 0,9, 1, 1)
CE, PQ, IQ	[0 1]	Extreme low	(0, 0, 0.1, 0.2)
		Very Low	(0.1, 0.2, 0.35)
		Low	(0.25, 0.40, 0.55)
		Moderate	(0.45, 0.55, 0.70)
		High	(0.60, 0.75, 0.9)
		Very high	(0,8 0,9, 1, 1)

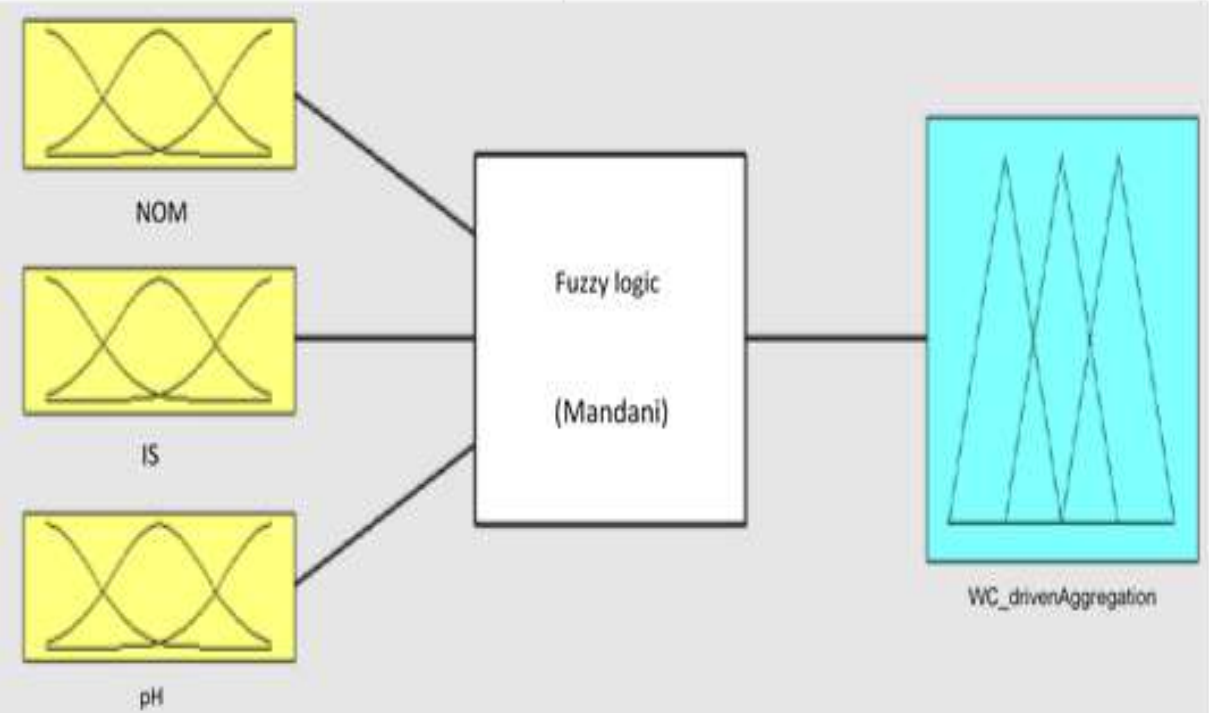


Figure D. 1. Mamdani fuzzy inference system (FIS) editor showing inputs for WCA

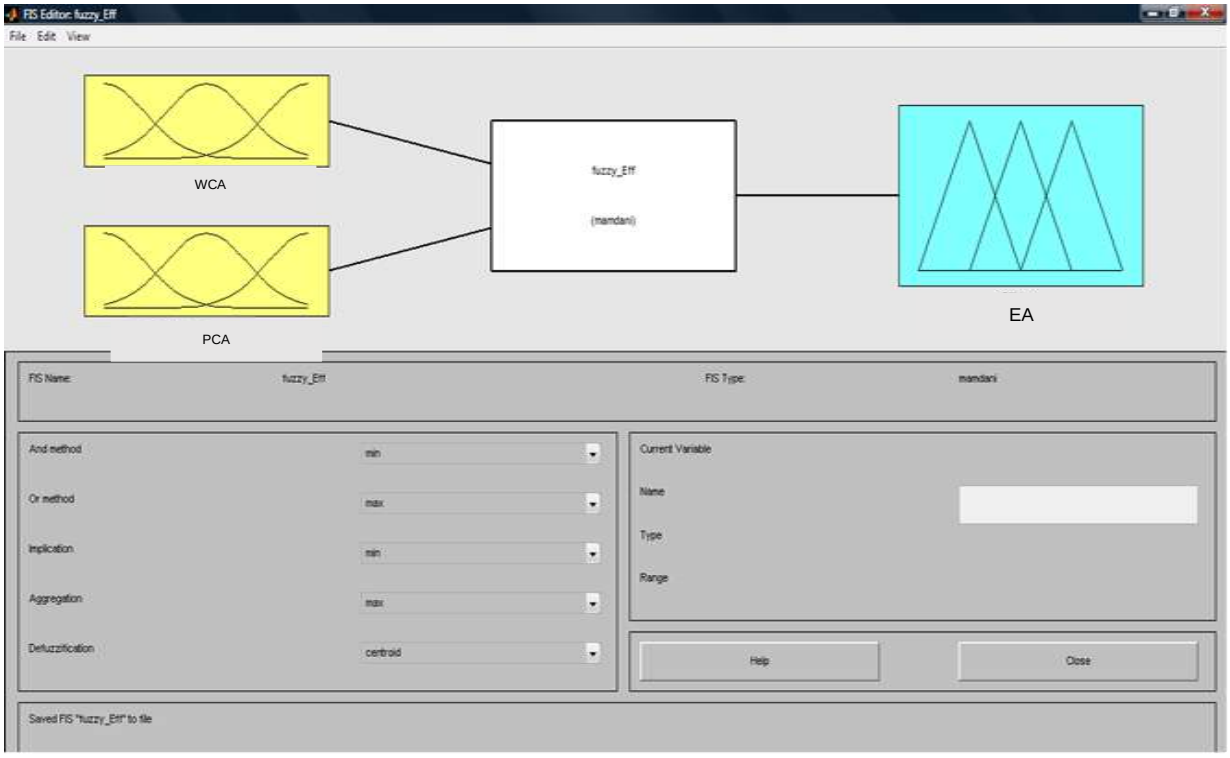


Figure D. 2. FIS editor for estimating the EA based on WCA and PCA as inputs.

Figure D. 3. Rule viewer for estimating the EA based on WCA and PCA as inputs.

Table D. 3. Fuzzy sets for the crisp inputs

Example No.	µTime (1)			µSolubility (2)			µEC (3)			µCoating (4)			µSize (5)			µZP (6)		
	Low	Moderate	High	Low	Moderate	High	Low	Moderate	High	Low	Moderate	High	Small	Moderate	Large	Low	Moderate	High
1Zn	1.00	0	0	0	0	1.00	1.00	0	0	1.00	0	0	0.50	0.33	0	0	0.31	0.69
2Zn	1.00	0	0	0	0	1.00	1.00	0	0	1.00	0	0	0.50	0.33	0	0	0.31	0.69
3Zn	1.00	0	0	0	0	1.00	1.00	0	0	0	0	1.00	0.50	0.33	0	0	0.31	0.69
4Zn	1.00	0	0	0	0	1.00	1.00	0	0	0	1.00	0	0.50	0.33	0	0	0.31	0.69
5Zn	1.00	0	0	0	0	1.00	1.00	0	0	1.00	0	0	0.50	0.33	0	0	0.31	0.69
6Zn	0	0	1.00	0	0	1.00	1.00	0	0	1.00	0	0	0.50	0.33	0	0	0.31	0.69

µNOM (7)			µpH (8)			µIS (9)		
low	Moderate	High	low	Moderate	High	Low	Moderate	High
0	0.90	0.05	1.00	0	0	0.78	0.15	0
0	0.90	0.05	1.00	0	0	0	0	1.00
1.00	0	0	1.00	0	0	0.78	0.15	0
1.00	0	0	1.00	0	0	0.78	0.15	0
1.00	0	0	0	0.75	0	0.78	0.15	0
1.00	0	0	0	0.75	0	0.78	0.15	0

Table D. 4. Fuzzy sets for the crisp inputs

Example No.	μTime (1)			μSolubility (2)			μEC (3)			μCoating (4)			μSize (5)			μZP (6)		
	Low	Moderate	High	Low	Moderate	High	Low	Moderate	High	Low	Moderate	High	Small	Moderate	Large	Low	Moderate	High
1Ti	1.00	0	0	1.00	0	0	1	0	0	1.00	0	0	0.50	0.33	0	0	0.32	0.69
2Ti	1.00	0	0	1.00	0	0	1	0	0	1.00	0	0	0.50	0.33	0	0	0.32	0.69
3Ti	1.00	0	0	1.00	0	0	0.70	0.33	0	0	0	1.00	0.50	0.33	0	0	0.32	0.69
4Ti	1.00	0	0	1.00	0	0	0.70	0.33	0	0	1.00	0	0.50	0.33	0	0	0.32	0.69
5Ti	1.00	0	0	1.00	0	0	0.97	0.04	0	1.00	0	0	0.50	0.33	0	0	0.32	0.69
6Ti	0	0	1.00	1.00	0	0	0.97	0.04	0	1.00	0	0	0.50	0.33	0	0	0.32	0.69

μNOM (7)			μpH (8)			μIS (9)		
Low	Moderate	High	Low	Moderate	High	Low	Moderate	High
0	0.90	0.05	0	0.50	0	0.78	0.15	0
0	0.90	0.05	0	0.50	0	0	0	1.00
1.00	0	0	0	0.50	0	0.78	0.15	0
1.00	0	0	0	0.50	0	0.78	0.15	0
1.00	0	0	1.00	0	0	0.78	0.15	0
1.00	0	0	1.00	0	0	0.78	0.15	0

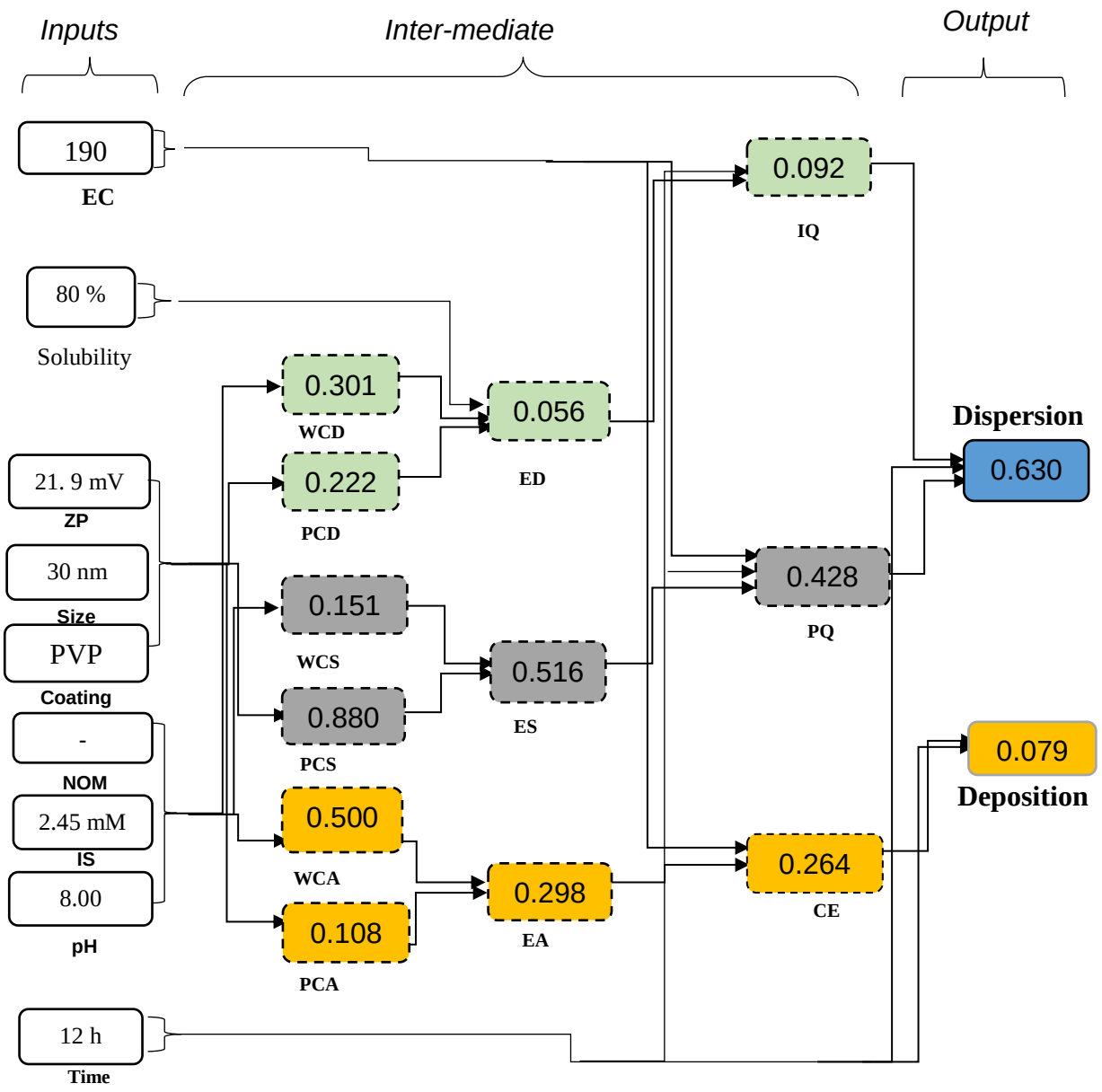


Figure D. 4. Illustrating stepwise functionality of FL using nZnO for Scenario 3

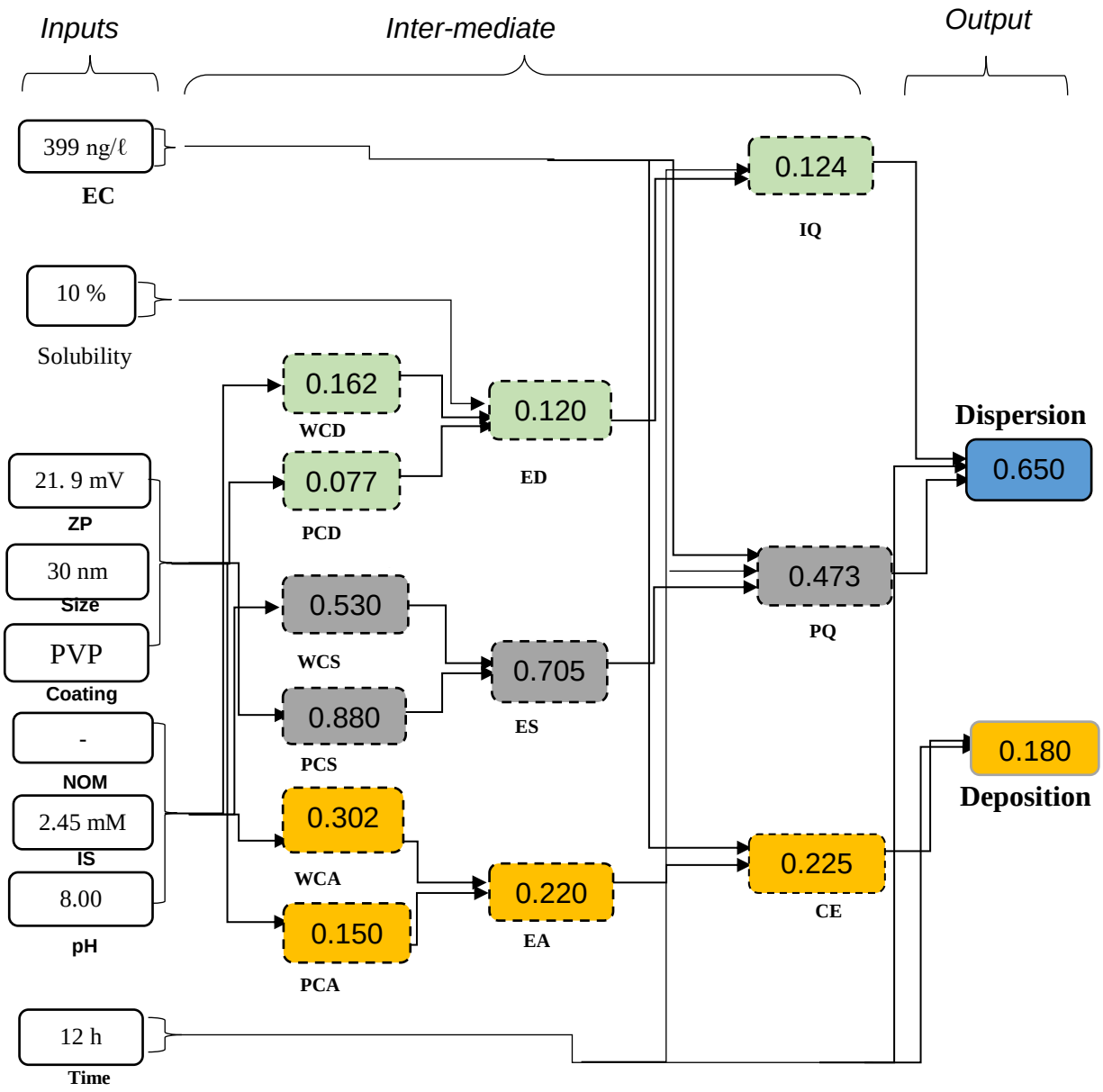


Figure D. 5. Illustrating stepwise functionality of FL using $nTiO_2$ for Scenario 3.

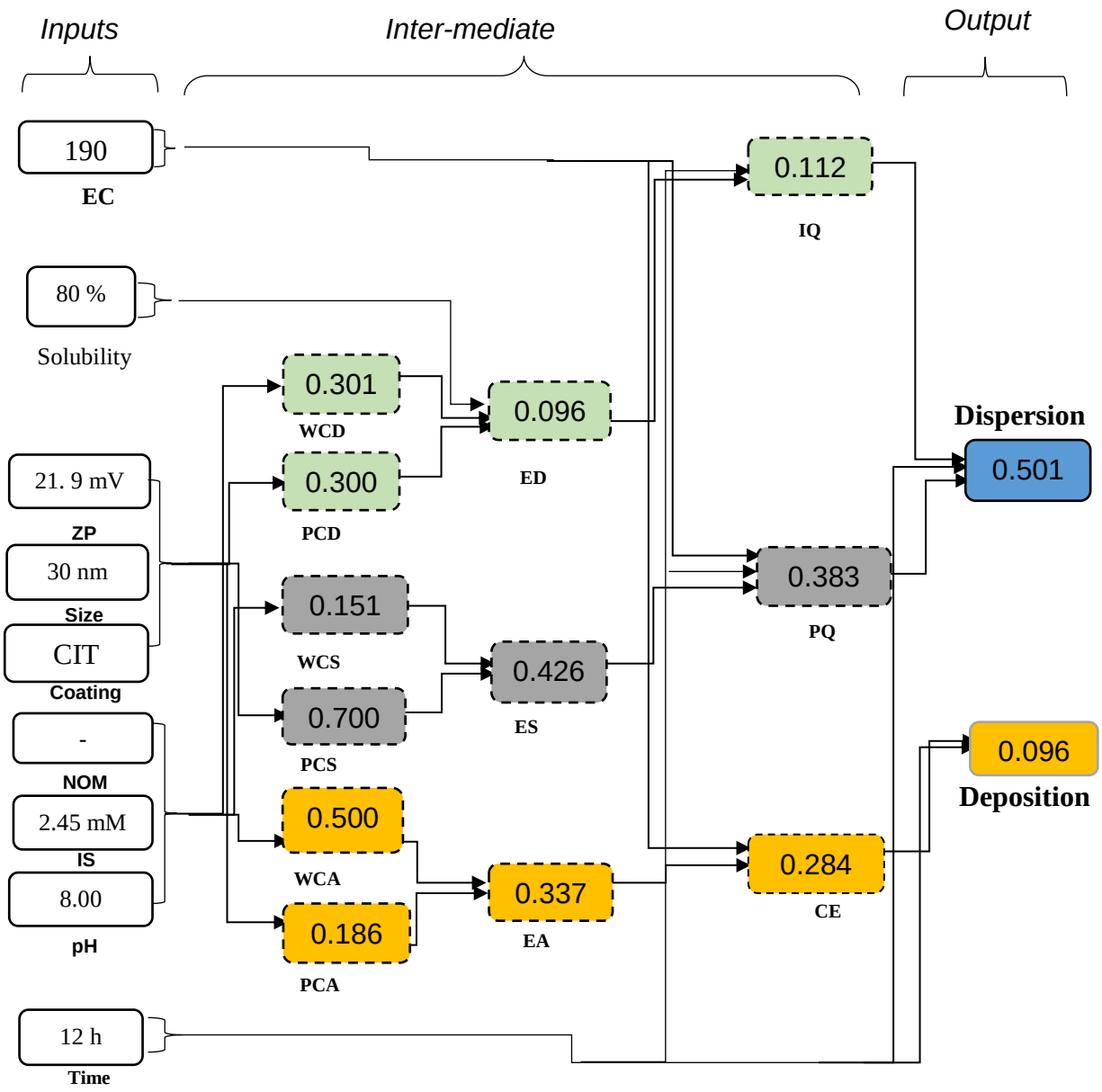


Figure D. 6. Illustrating stepwise functionality of FL using nZnO for Scenario 4.

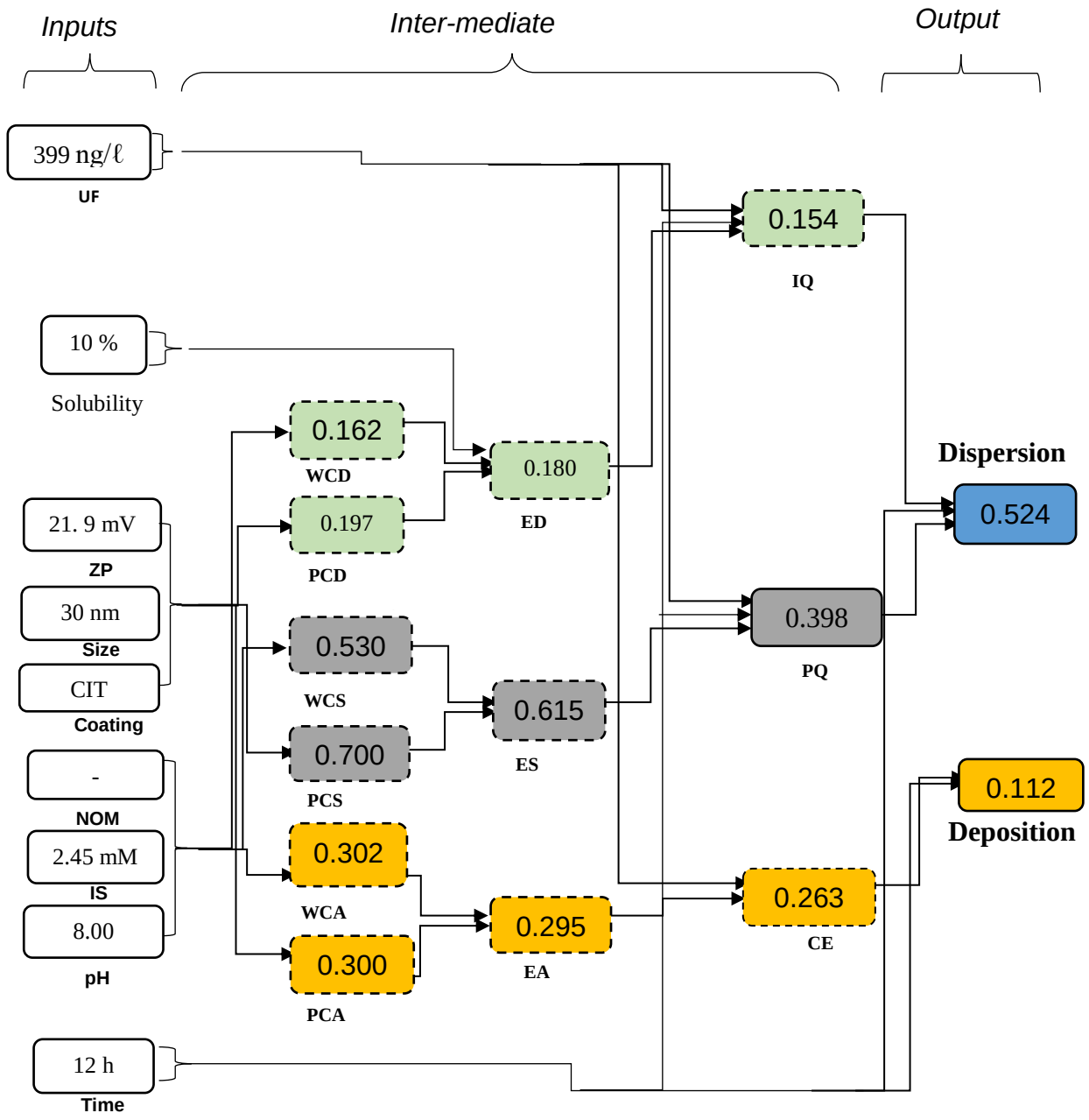


Figure D. 7. Illustrating stepwise functionality of FL using $nTiO_2$ for Scenario 4

.

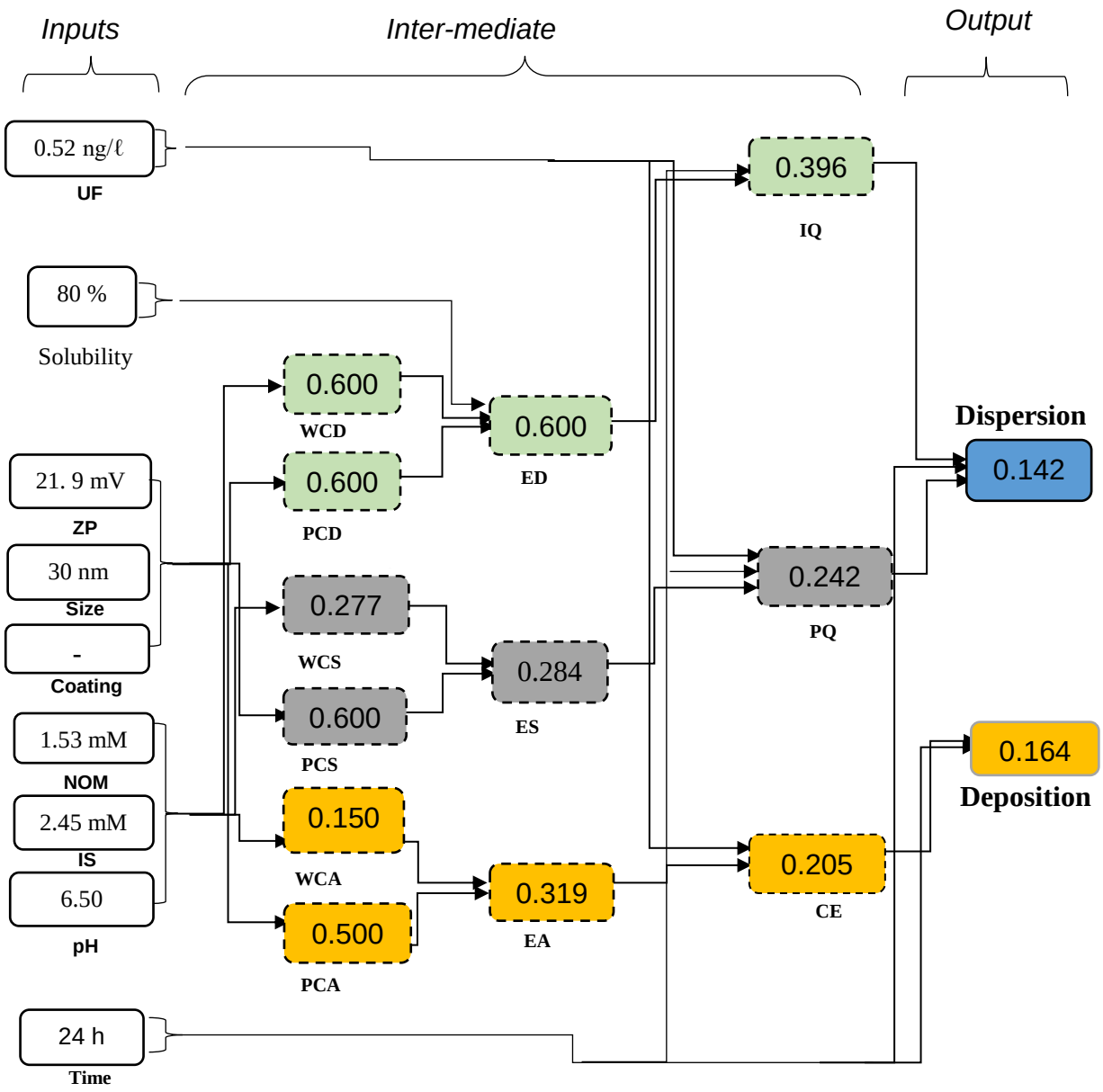


Figure D. 8. Illustrating stepwise functionality of FL using nZnO for Scenario 5

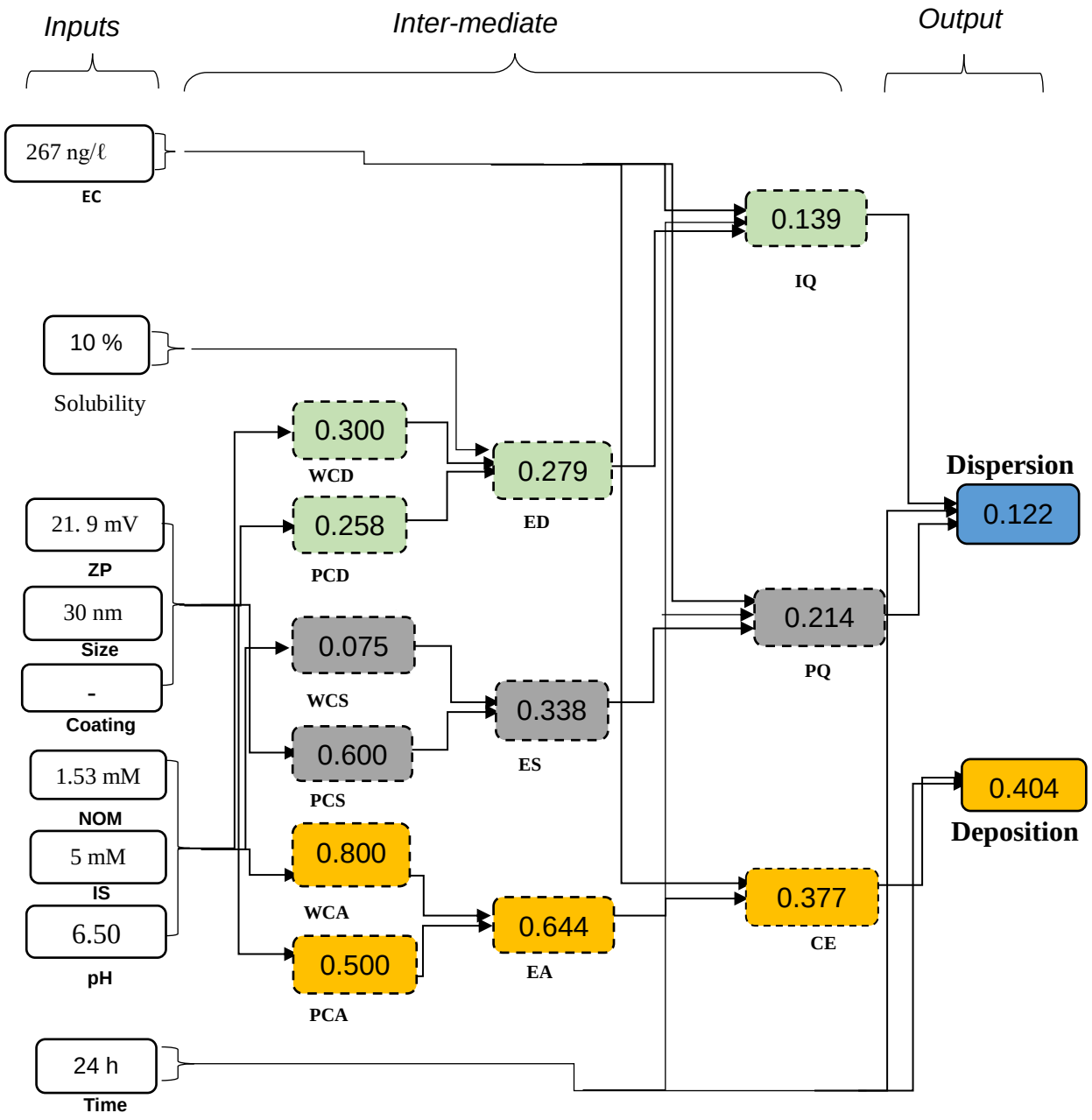
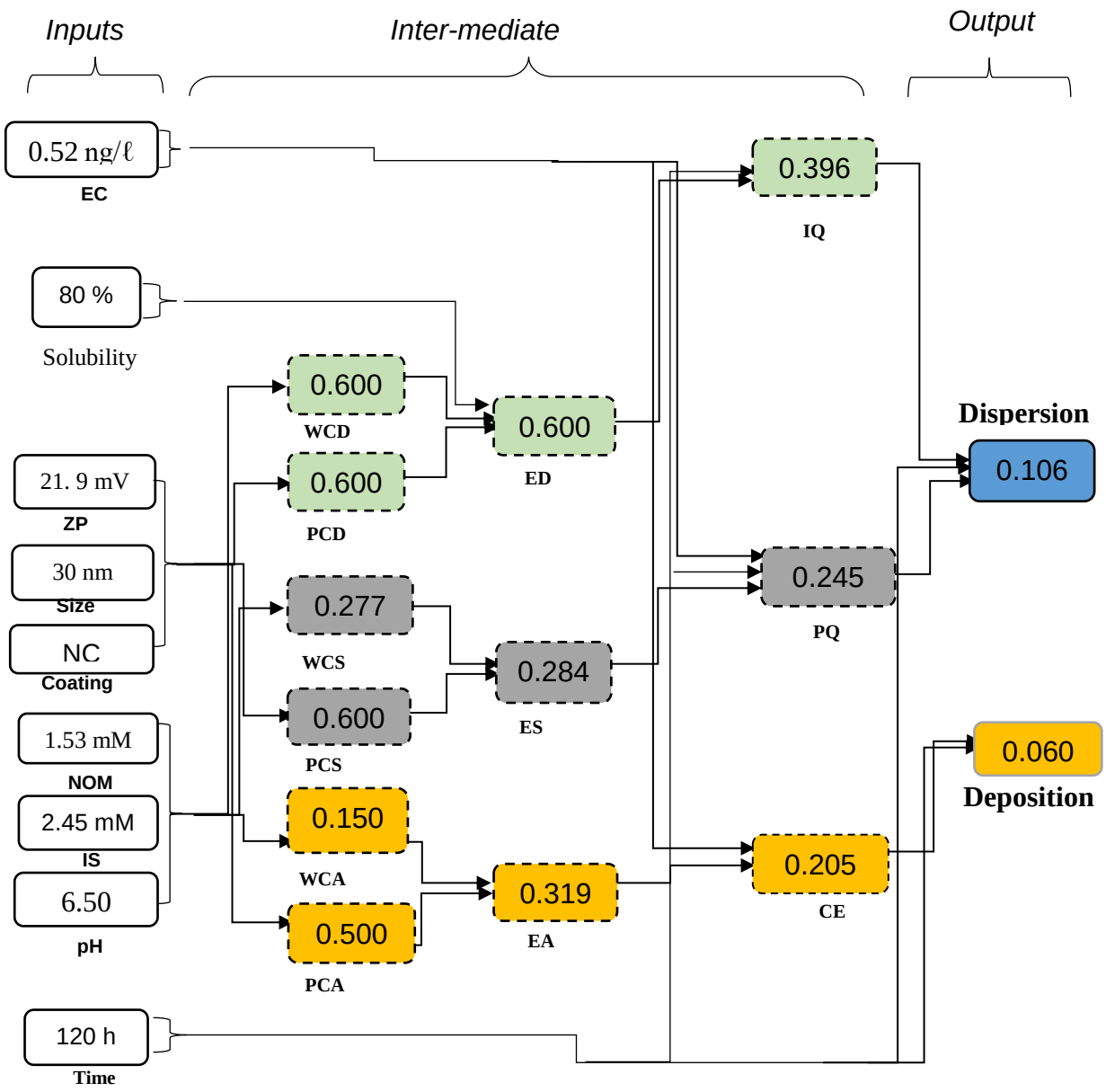


Figure D. 9. Illustrating stepwise functionality of FL using $nTiO_2$ for Scenario 5

1 .



2 Figure D. 10. Illustrating stepwise functionality of FL using nZnO for Scenario 6.

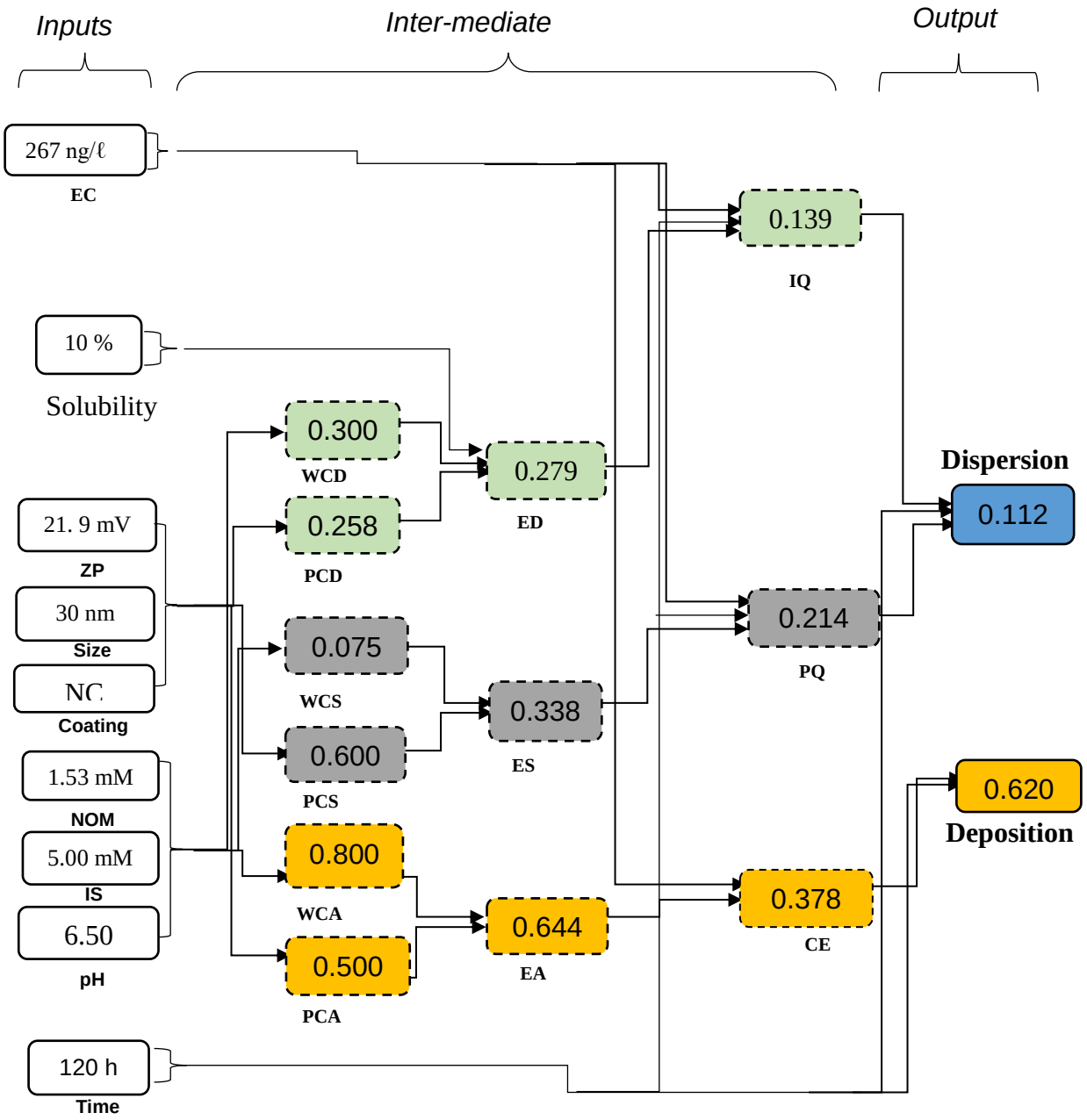
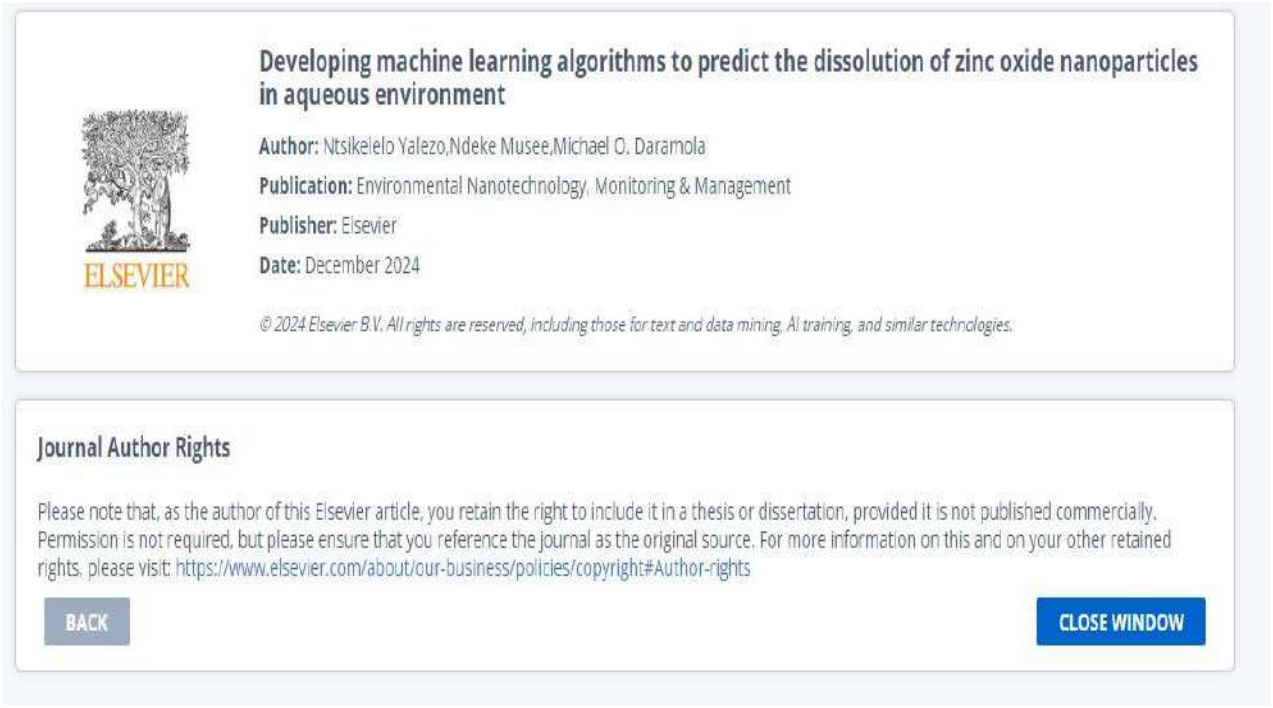


Figure D. 11. Illustrating stepwise functionality of FL using $nTiO_2$ for Scenario 6

10

Copyright permission



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

Author: Ntsikelelo Yalezo,Ndeke Musee,Michael O. Daramola

Publication: Environmental Nanotechnology, Monitoring & Management

Publisher: Elsevier

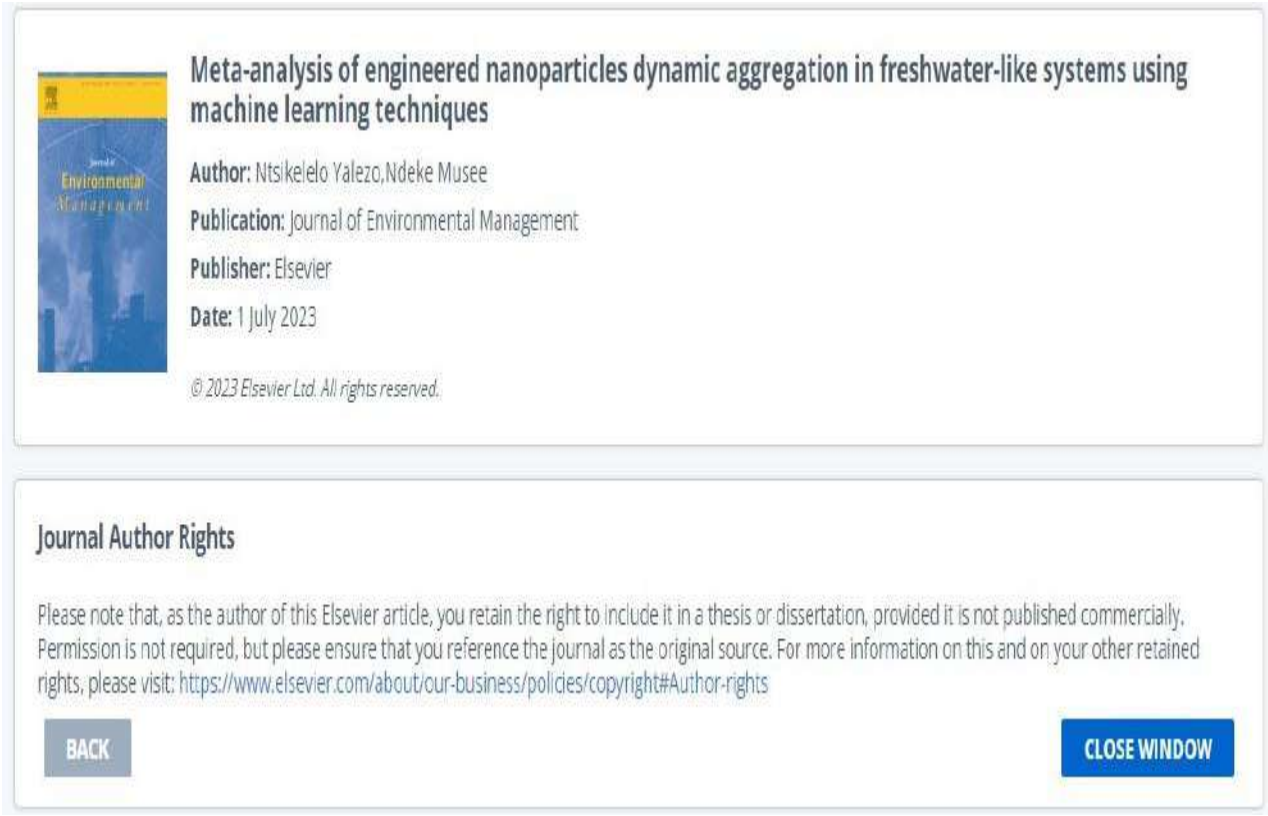
Date: December 2024

Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

BACK CLOSE WINDOW

11



Meta-analysis of engineered nanoparticles dynamic aggregation in freshwater-like systems using machine learning techniques

Author: Ntsikelelo Yalezo,Ndeke Musee

Publication: Journal of Environmental Management

Publisher: Elsevier

Date: 1 July 2023

Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

BACK CLOSE WINDOW

12

13