Environmental Advances 17 (2024) 100554

Contents lists available at ScienceDirect

Environmental Advances

journal homepage: www.sciencedirect.com/journal/environmental-advances

Artificial intelligence and machine learning in environmental impact prediction for soil pollution management – case for EIA process

Babatunde Anifowose [a,b,*], Fatai Anifowose [c]

[a] College of Engineering, Environment & Science, Coventry University, Coventry CV1 5FB, UK
[b] Centre for Agroecology, Water and Resilience (CAWR), Coventry University, Wolston Ln, Ryton-on-Dunsmore, Coventry CV8 3LG, UK
[c] Saudi Arabian Oil Company, Dhahran 31311, Saudi Arabia

ARTICLE INFO

Keywords:
Machine Learning
Artificial Intelligence
Multivariate Linear Regression
Sustainability
Environmental Impact Assessment
Soil Quality
Environmental Data Science

ABSTRACT

Scientific predictions are a key component of Environmental Impact Assessments (EIA), which can indicate the level of change within an environmental sphere (e.g., soil). As part of the EIA process, decision-making in mitigating complex environmental problems such as maintaining soil quality can be challenging, especially in data-sparse locations. Artificial Intelligence (AI) can ameliorate but the literature suggests that the deployment of Machine Learning (ML) techniques in soil research is concentrated mostly in developed countries. The potential of ML in managing soil pollution from complex mixture of heavy metals, petroleum hydrocarbons, and physicochemical factors is rarely explored. To address this research gap, we built robust models that increase the accuracy of impact prediction based on new experimental soil data from a data-sparse region of Africa (i.e., Nigeria). The algorithms applied are artificial neural networks (ANN), support vector regression (SVR), regression tree (RT), and random forest (RF). The study also implemented a multivariate linear regression (MLR) model as a baseline. Key findings include (a) the MLR model performed less than the machine learning models largely due to the nonlinearity of data; (b) Log-normalization helped to improve the predictive capability of all models as the effects of statistical variability were removed; (c) the RF model had the best performance in terms of correlation coefficient, mean absolute error, and root mean square error, and (d) the machine learning models showed improved performance with increased correlation and lower error between the actual and predicted soil electrical conductivity values. Our results imply that data sparsity may no longer be an excuse for the non-use of quantitative impact prediction in Environmental Impact Assessment (EIA) processes. This could change how EIAs are conducted and enhance sustainability in natural resource exploitation, globally. Future work will apply algorithms for automated feature selection to obtain optimal subset of soil quality measurements that will further improve the accuracy of the models.

1. Introduction

Environmental Impact Assessment (EIA) is a "proactive methodical process that investigates and predicts the potential direct, indirect and cumulative impacts of proposed project activities on environmental receptors, ideally from project initiation to decommissioning, and offers mitigation strategies" (Anifowose et al. 2016, p.571). A typical environmental receptor in EIAs is soil; and soils are critical to the sustainability of global and local environments (Komatsuzaki and Ohta 2007; Nosova and Uspenskaya 2023). Yet, soil degradation is mostly associated with the manufacturing, production, utility and energy industries, which remains a major problem globally (George et al. 2021; Varjani and Upasani 2019). Heavy metals such as Lead (Pb), Zinc (Zn), Copper (Cu), Iron (Fe), Nickel (Ni), Chromium (Cr), and Vanadium (V) and physicochemical parameters including pH and electrical conductivity (EC) are extensively studied in soils (Kazemi and Hosseini 2011; Teng et al. 2014; Li et al. 2019; Yang et al. 2020). Also, Polycyclic Aromatic Hydrocarbons (PAHs) and Total Petroleum Hydrocarbons (TPHs) are often the focus of oil pollution related studies (Pascoe et al. 1998; Douglas et al. 2004; Ugochukwu et al. 2018; Akinpelu et al. 2020). The inclusion of compounds such as PAHs and volatile hydrocarbons like Benzene, Toluene, Ethylbenzene and Xylenes (BTEX) can improve the soil quality assessment process (Pinedo et al. 2013) while TPH is useful in predicting risk parameters for petroleum products (Ugochukwu et al. 2018).

* Corresponding author.
E-mail addresses: b.anifowose@coventry.ac.uk, tundean@yahoo.com (B. Anifowose).

https://doi.org/10.1016/j.envadv.2024.100554
Received 2 January 2024; Received in revised form 29 May 2024; Accepted 30 May 2024
Available online 4 June 2024

Nomenclature

pH	potential of Hydrogen
Cl	Chlorine
Na	Sodium
k	Potassium
TPH	Total Petroleum Hydrocarbons
TAH	Total Aliphatic Hydrocarbons
PAH	Polycyclic Aromatic Hydrocarbons
BTEX	Benzene, Toluene, Ethylbenzene and Xylenes
Fe	Iron
EC	Electrical Conductivity
NCP	National Council on Privatization
ML	Machine Learning
AI	Artificial Intelligence
ANN	Artificial neural networks
SVR	Support Vector Regression
RT	Regression Tree
RF	Random Forest
MLR	Multivariate Linear Regression
EIA	Environmental Impact Assessment
PPMC	Products Marketing Company Limited
GPS	Global Positioning System
BPE	Bureau of Public Enterprise

Meanwhile, machine learning (ML), the computational aspect of Artificial Intelligence (AI), is known to simplify complex systems, and reduce computational burden and time (Wu 2019; Brunswick et al. 2021). For example, Anifowose and Abdulraheem (2011) blended three ML techniques to predict oil and gas reservoir properties (porosity and permeability) and concluded that hybrid models tend to yield better quality information and predictive accuracy (Chou et al. 2021) than individual techniques as in Bieganowski et al. (2018). Also, ML algorithms consistently yielded more accurate results in comparison with empirical models in the prediction of leaf wetness as complex environmental conditions complicated the models (Gillespie et al. 2021). Multi-label data frame for model training in deep learning-based ML algorithms also tend to perform better (Matsui et al. 2022) just as Anggraini et al. (2024) demonstrated the usefulness of multiple models in enhancing accuracy than single models. To reduce uncertainties and ensure best solutions, models that address multiple hypotheses can evolve through a novel paradigm called ensemble machine learning as expatiated in Anifowose et al. (2017). Data is the medium through which various phenomena communicate their conditions with their environment. Hence, accurate prediction can help monitor and control resource use to mitigate adverse environmental impacts often detailed in EIAs, and ML algorithms offer such opportunities (Song et al. 2020).

A number of state-of-the-art techniques such as support vector regression (SVR), random forest (RF) and extreme learning machines are still waiting to fully benefit the environmental modelling processes, although optimization algorithms have recently been combined with support vector machine (SVM) in environmental governance engineering (Ai and Yang, 2019; da Silveira et al. 2022). Several other techniques leverage the advances in hybrid and ensemble learning methodologies to improve the predictive capabilities of current models. For instance, Yang et al. (2017) employed RF, artificial neutral network (ANN) and SVR in a study of reservoir inflows in the USA and China and their results showed that RF had the best statistical performance in comparison with ANN and SVR while Aftab et al. (2022) found ANN and SVR models to be highly accurate. Similarly, Hou et al. (2020) found RF model to have the best predictive performance just as Wu et al. (2019) posit that RF is "one of the best classification algorithms".

1.1. Problem Statement, Motivation, and Scope of Study

Studies have suggested that ML application "ignores soil science knowledge" including the possibility that results are misleading and wrong (Rossiter 2018; Padarian et al. 2020). However, the application of ML in soil science research has historically focused on developed countries (mostly in temperate regions) partly due to the digital divide in most developing countries (Padarian et al. 2020). Forkuor et al. (2017) confirm this by highlighting the sparsity of studies that compare traditional regression and different ML approaches to predict soil properties in West Africa. The limited application of ML in soil science in developing nations has been linked to the correlation between science, technology and development by Padarian et al. (2020). In their recent study, they found that only approximately 20 % of Africa has institutional affiliations in a Google Scholar search completed on 1 February 2019, which returned more than 70,000 articles using the keywords "soil" and "machine learning". Meanwhile, the claim that ML "ignores soil science knowledge" and probably yields misleading/wrong results can only be acceptable or debunkable when fully tested across the world. Hence, one of our study objectives is to use information from this under-represented part of the world to assess how ML algorithms autonomously handle nonlinearity in soil data. This is important because soil properties across temperate and tropical regions are not always similar (Eijsackers et al. 2017) just as their abilities for the uptake of pollutants differ e.g. temperate soils experience twice as slow carbon turnover than tropical soils (Six et al. 2002).

Hengl et al. (2017) applied two ML algorithms to spatially predict 15 soil macro and micronutrients across Sub-Saharan Africa. The results revealed that manganese, aluminium, zinc, boron and sodium are the important nutrients for predicting crop yield. Papageorgiou et al. (2009) also linked cotton production to pH, phosphorus, potassium, calcium, nitrogen, and magnesium using fuzzy cognitive maps. Forkuor et al. (2017) used remote sensing to map six soil properties viz: sand, silt, clay, cation exchange capacity, soil organic carbon and nitrogen while comparing four ML models. According to Anifowose et al. (2014), identification of oil pollution location along linear facilities can be a major problem in some areas, but Ozigis et al. (2019) deployed ML's random forest classifier using Landsat 8 (OLI spectral bands) and Vegetation Health Indices to identify oil-impacted areas. All these ML studies, thus far, tend to focus mainly on spatiotemporal mapping for agricultural yields and soil features. Clearly, the application of ML techniques in the prediction of soil sediment pollution in which heavy metals, petroleum hydrocarbons, and selected physicochemical parameters are combined for holistic appraisal of soil quality is rare. To the best of our knowledge, no such study is found in the literature nor focusing on the tropics i.e., a data sparse region. This deserves research attention given the renewed calls for environmental sustainability as novel and efficient approaches to environmental impact prediction are now more essential. In addition, there is a gap in our understanding of the efficacy of ML techniques in managing soil pollution occasioned by complex mixture of heavy metals (e.g. Pb, Zn, Cu), petroleum hydrocarbons (e.g. TPH, BTEX) and physicochemical factors (e.g. electrical conductivity). To address this gap, we present a modest first effort to achieve the following objectives:

(a) demonstrate that the complex mixture of soil sediment data is nonlinear; and hence, cannot be adequately analysed using the traditional empirical approaches and conventional statistical linear regression method, as may have been the case in previous EIAs and/or similar impact assessment reports;
(b) examine the robustness and efficiency of ML techniques in handling such nonlinear problems, where heavy metals and

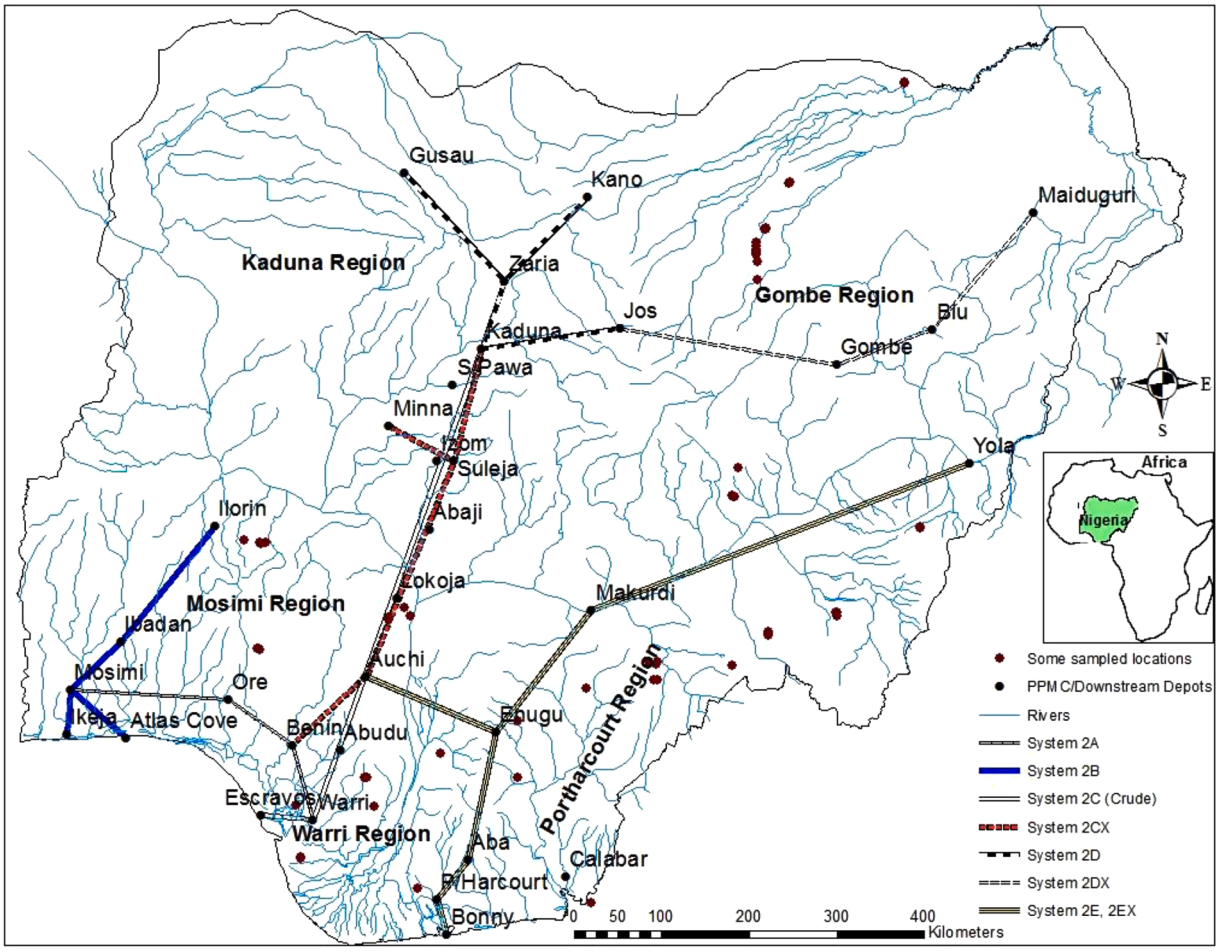

Fig. 1. Nigeria showing the PPMC oil facilities and the sampling locations for the Environmental Audit study. Inset: Map showing Nigeria in West Africa. Source: Anifowose and Odubela (2018)

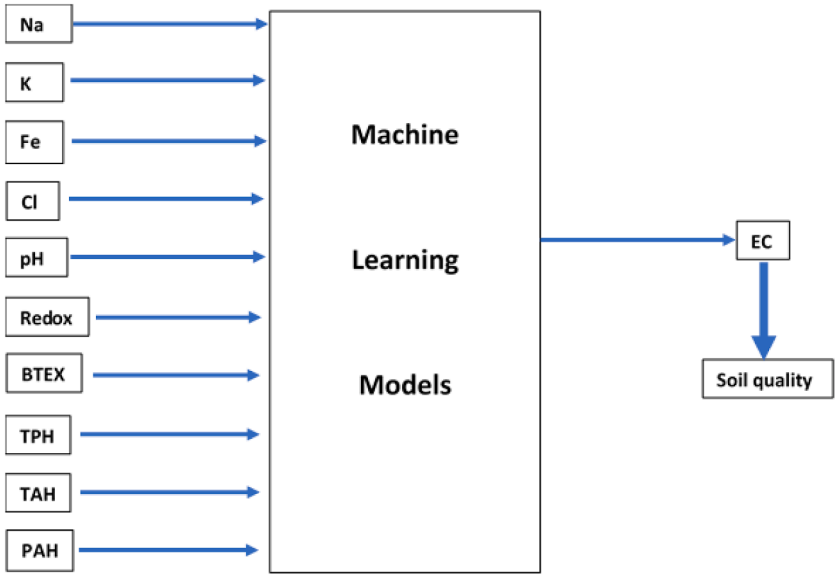

Fig. 2. An illustrative diagram of input data and methods applied in this study

petroleum hydrocarbons are found in soil samples, by modelling electrical conductivity in soil; and,

(c) further assess, in a tropical region context, the hypothesis that ML "ignores soil science knowledge" and that its results might be misleading or wrong.

Table 1
Statistical Description of the dataset

	pH	Redox	Cl	Na	K	TPH	TAH	PAH	BTEX	Fe	EC
Min	0.69	0.605	28	9.46	1.91	0.9284	0.0024	0	0	0.89	0.029
Max	8.5	6389.53	2710	96014	131.03	251543	116400	78625	194.24	9537.5	2392.39
Range	7.81	6388.92	2682	96004.54	129.12	251542.07	116399.99	78625	194.24	9536.61	2392.36
STD	0.74	932.36	329.87	7173.26	18.92	37323.27	18868.88	9747.25	32.91	787.25	383.83
VAR	0.55	869292.74	108812.04	51455604.89	358.16	1393026376	356034532.4	95008955.62	1083.34	619770.62	147326.38

NB: pH – power of Hydrogen; Cl – Chlorine; Na – Sodium; k – Potassium; TPH – Total Petroleum Hydrocarbons; TAH – Total Aliphatic Hydrocarbons; PAH – Polycyclic Aromatic Hydrocarbons; BTEX – Benzene, Toluene, Ethylbenzene and Xylenes; Fe – Iron; EC – Electrical Conductivity.

Table 2
Correlation between EC, other physiochemical parameters and petroleum hydrocarbons

	pH @ 250C	**Redox (mV)**	**Cl^- (mg/l)**	**Na (mg/l)**	**K (mg/l)**	**TPH**	**TAH**	**PAH**	**BTEX**	**Iron (mg/l)**	**EC (Us/cm)**
pH @ 25^{0}C	1.000										
Redox (mV)	-0.099	1.000									
Cl (mg/l)	0.088	0.044	1.000								
Na (mg/l)	0.021	-0.047	-0.023	1.000							
K (mg/l)	0.012	-0.073	-0.112	0.078	1.000						
TPH	-0.092	0.449	-0.032	-0.022	0.064	1.000					
TAH	-0.129	0.539	-0.027	-0.023	0.035	0.884	1.000				
PAH	-0.025	0.484	0.032	-0.020	0.037	0.625	0.709	1.000			
BTEX	-0.014	0.328	-0.038	-0.037	-0.145	0.269	0.229	0.170	1.000		
Iron (mg/l)	0.042	0.160	0.215	-0.039	-0.085	0.107	0.105	0.028	0.279	1.000	
EC (Us/cm)	**-0.028**	**-0.178**	**0.208**	**-0.020**	**-0.052**	**-0.074**	**-0.079**	**-0.068**	**-0.100**	**-0.085**	**1.000**

NB: pH – power of Hydrogen; Cl – Chlorine; Na – Sodium; k – Potassium; TPH – Total Petroleum Hydrocarbons; TAH – Total Aliphatic Hydrocarbons; PAH – Polycyclic Aromatic Hydrocarbons; BTEX – Benzene, Toluene, Ethylbenzene and Xylenes; Fe – Iron; EC – Electrical Conductivity.

Table 3
Results of the Initial Run of Models on Raw Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations.

Models	R-Square		P-Value		RMSE	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**
MLR	0.45	0.01	0	0.015	333.81	457.97
ANN	0.71	0.70	0.0	0.0	296	304.86
SVR	0.15	0.10	0.07	0.47	386.36	384.76
RT	0.77	0.74	0.0	0.0	240.45	273.10
RF	0.80	0.86	0.0	0.0	261.63	218.90

NB: Multivariate Linear Regression (MLR); Artificial Neural Networks (ANN), Support Vector Regression (SVR), Regression Tree (RT); Random Forest (RF).

2. Data and Methodology

2.1. Data Sources and Soil Parameters

It is not uncommon to find data paucity, demonstrated through incomplete, not up-to-date, and limited monitored environmental parameters in developing nations (Le Goff et al. 2022) and sometimes in developed nations (Nosova and Uspenskaya 2023; Anggraini et al. 2024). The data used for this study are based on an environmental audit survey of the Pipelines and Products Marketing Company (PPMC) of Nigeria in West Africa (Fig. 1). More details on the study area and the related facilities can be found in Anifowose et al. (2012, 2014) and Anifowose and Odubela (2018). The Federal Government of Nigeria through the National Council on Privatization (NCP) and the Bureau of Public Enterprises (BPE) commissioned the environmental audit of the PPMC facilities and was co-funded by the World Bank. The data utilised in this present article were retrieved from the PPMC environmental audit data archive and report i.e. NCP/BPE (2008).

Soil EC is a function of geotechnical properties such as soil temperature, degree of water saturation, pore water salinity amongst others (Alsharari et al. 2020). EC is one of the principal indicators of soil quality, which has revealed underlying factors influencing soil properties under different conditions (Mao et al. 2016; Stenchly et al. 2017; Kim et al. 2018; Zhang et al. 2020). EC measures the amount of dissolved material (e.g. salts, impurities, minerals etc.) in a given soil sample. Studies have shown an inverse relationship between EC and petroleum hydrocarbons in contaminated soils, where for example, a lower hydrocarbon content in fresh or biodegraded oil corresponds to a higher conductivity (Mao et al. 2016). TPH is a pollutant – an impurity – that can be found in oil-polluted soil; and TPH is often the hydrocarbon parameter used for examining soil electrical conductivity variations (e.g. Mao et al. 2016). Abdel Aal and Atekwana (2014) found that elevated electrical conductivity was one of two factors partly related to the biogeochemical changes in oil composition; and often one of the monitored parameters in water quality studies (Le Goff 2022).

Therefore, this study attempts to model EC measurements from soil sediment samples based on TPH, BTEX, PAH and TAH on one hand and a number of heavy metals on the other as illustrated in Fig. 2. The decision was made to include as much soil quality indicators as possible to leverage the capability of ML methods to "improve their performance automatically through experience" (Kim and Park 2009). These algorithms can identify previously unknown relationships by modelling nonlinear correlations and interactions (Chen et al. 2019) whilst overcoming the inadequacy in expert knowledge (e.g., Papageorgiou et al. 2009). This is a potentially key benefit for the future of EIA processes, especially as the laboratory quantification of some hydrocarbon components like PAH is far more difficult than others while the measurement of soil EC is less expensive, much easier and quicker (e.g. Seifi and Alimardani 2010). Therefore, this is not a mere soil pollutants mapping study but one that demonstrates how ML techniques could enhance impact prediction in future EIAs.

2.2. Sample preparation and data collection

A combination of physical observations and geo-referenced sampling of facility locations using handheld Global Positioning System (GPS) receivers were employed for the data collection process. Following a reconnaissance survey of the facilities, soil sampling and land use investigations along transects and within a 0.5 km radius of depots as well as along orthogonal transects at 0.5 km on both sides of pipelines were undertaken (NCP/BPE 2008).

At sampling locations, soil samples were collected from product

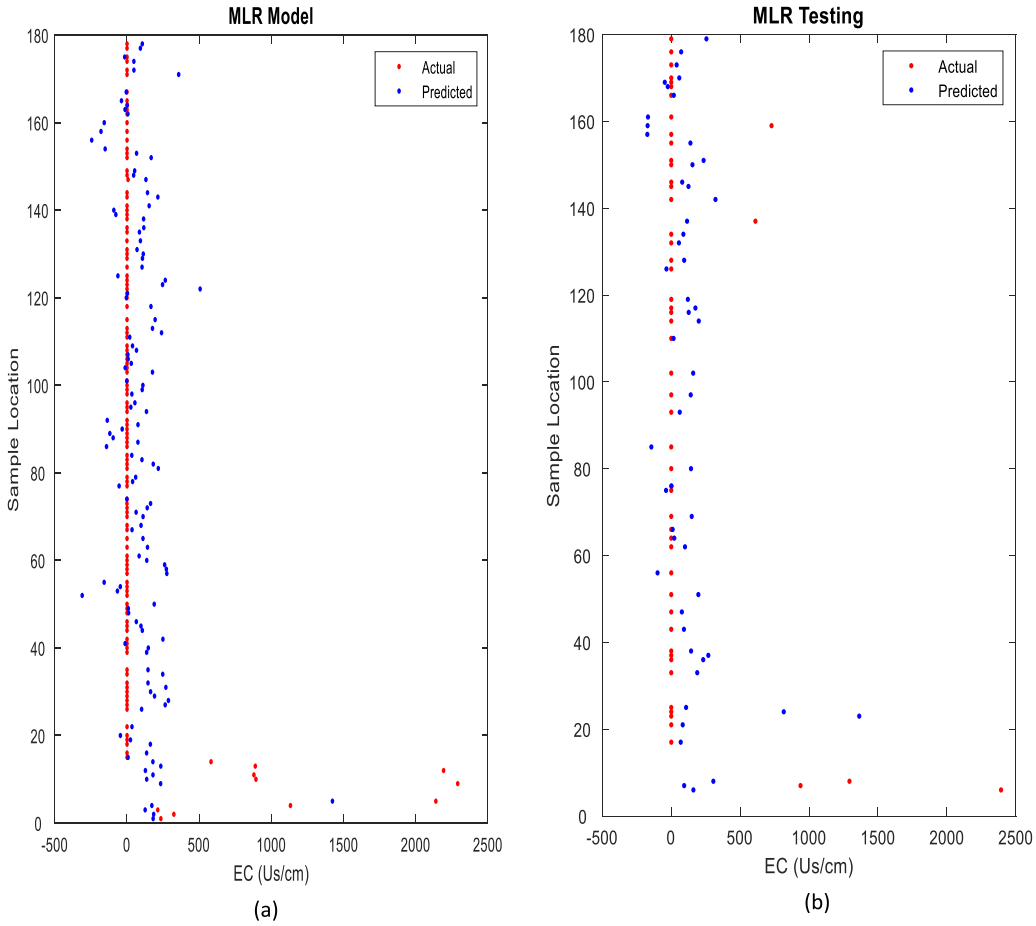

Fig. 3. Actual and Predicted (a) Training and (b) Testing Results of the MLR Model on Raw Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations. NB: The training and testing data were randomly selected using a stratified sampling approach where 70 % of the data was selected for training and 30 % reserved for blind test.

storage facilities (tanks), transfer and export facilities, power generators, fire and service water systems, skimming pit/oil trap system, drain systems, product reception areas, and staff clinics (NCP/BPE 2008). Along the pipelines, sampling was carried out in ecologically sensitive areas like wetlands, past oil affected areas, and specific geomorphic/vegetation units. Soil auger was used for sampling at each sampling points from 0-15 cm, 15-50 cm and 50-100 cm soil depth; similar to the 0-20 cm approach by Wang et al. (2019) and the 0-10 cm by Yang et al. (2020).

According to NCP/BPE (2008), the collected soil samples were packed and preserved for onward transmission to the laboratory for subsequent physicochemical analyses. The data include heavy metals such as Fe, Cd, Cr, Pb, Cu, Ni, V, and Zn (Aftab et al. 2022). The policy procedures of the Federal Environmental Protection Agency (FEPA, 1991) and that of the Department of Petroleum Resources (DPR, 2002) guided the sampling methods and measurements. A total of 730 soil samples were analysed for the key pollution indicators broadly categorised as heavy metals, petroleum hydrocarbons and physicochemical parameters in the report (NCP/BPE 2008). Only 179 of these soil samples was found useful for this study. This is majorly due to incomplete measurements characterized by blank entries, and recording errors such as having character entries where numeric values are expected. This basically reveals the challenge faced by practitioners on data availability and quality in this field. For example, Wang et al. (2019) found only 2 % of data points (equivalent to only 74 sites) useful from the tropical regions out of approximately 52,000. Similar study on heavy metal distribution in sediments under the Caspian Ecosystem Program used only 80 surface sediment samples across five countries (Kazemi and Hosseini 2011).

2.3. Measurement of electrical conductivity (EC)

To measure EC at sampling locations and pipeline routes shown in Fig. 1, the Schlumberger array tool with an ABEM SAS 300C Tetrameter was used for surface geographical surveys in form of direct current electrical resistivity analysis (NCP/BPE 2008). The soil resistivity at 150 – 200 cm depth interval was determined in line with the usual depth of pipeline burial, which is 0.9 to 1 m. The resistance (R) of the subsurface was measured at an electrical current strength of 20mA. A geometric factor (K) was calculated for each spread and combined with the resistance to calculate the bulk apparent resistivity (ρ_a) as recommended in NCP/BPE 2008 (p.109) as shown in eq. (1).

$$\rho_a = \pi (l^2 - r^2) R/2\pi r \quad (1)$$

where ρ_a is the apparent resistivity (ohm m), l is the half current electrode spacing (m), r is the half potential electrode spacing (m), and R is the Resistance (ohms). Further details on soil electrical resistivity can be found in Alsharari et al. (2020).

NCP/BPE (2008: Chapter 4, p.68) states “the value of AB/2 (where AB is the inter-electrode spacing) was plotted against the corresponding ρ_a values on a logarithmic graph using the computer software **IPI2WIN** version 2.1”. Using the log-log graph that had the curve-of-best-fit and an overlaid EQUIVALENCE plot, the different geo-electric sections were established and this permitted the creation of corresponding lithologic inferences and depth delineations (NCP/BPE 2008). Since soil resistivity is the reciprocal of conductivity, it is often admissible as the primary

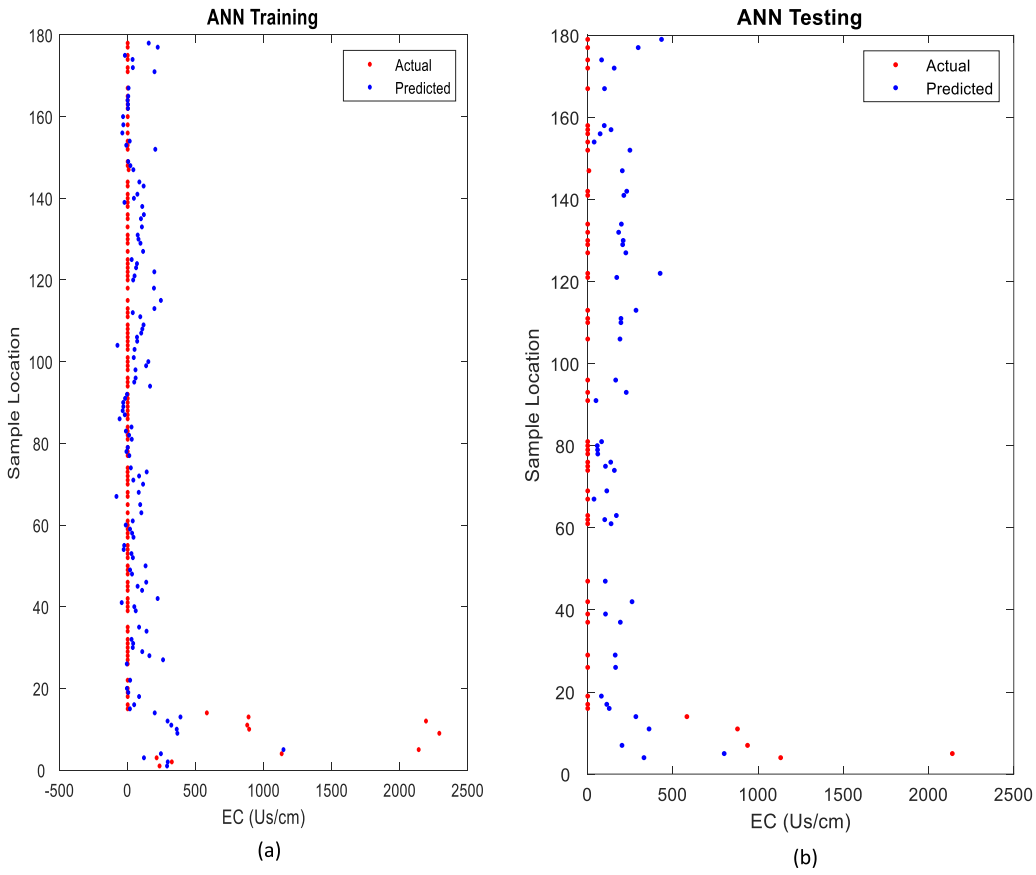

Fig. 4. Actual and Predicted (a) Training and (b) Testing Results of the ANN Model on Raw Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations. NB: The training and testing data were randomly selected using a stratified sampling approach where 70 % of the data was selected for training and 30 % reserved for blind test.

Table 4
Results of the second run of models on log-normalized measurements of heavy metals, petroleum hydrocarbons and ec from sampled oil facility locations.

Models	R-Square		P-Value		RMSE	
	Training	Testing	Training	Testing	Training	Testing
MLR	0.84	0.75	0	0	0.49	0.61
ANN	0.94	0.85	0	0	0.30	0.49
SVR	0.80	0.79	0	0	0.90	0.93
RT	0.85	0.89	0	0	0.48	0.45
RF	0.91	0.86	0	0	0.45	0.60

NB: Multivariate Linear Regression (MLR); Artificial Neural Networks (ANN), Support Vector Regression (SVR), Regression Tree (RT); Random Forest (RF).

indicator of soil corrosivity (Cunat 2002; Pritchard et al. 2013; Alsharari et al. 2020). A lower resistivity makes it easier for current to flow through the soil, thereby increasing soil corrosivity; and the apparent resistivity is calculated for the spacing between the pairs of electrodes using eq. (1).

Anifowose and Odubela (2018) and NCP/BPE (2008) contain a robust detail of the Quality Assurance/Quality Control (QA/QC) framework and analytical procedures for the sampling and data collection processes, including the recommended test methods for physico-chemical parameters' analyses as used in this study. In addition, the guidelines provided by the DPR and the FME helped to ensure the standardisation of data collection processes from the sites.

2.4. Proposed methods of modelling

The study implemented five algorithms for this study. One is the traditional multivariate linear regression (MLR) approach used as a benchmark to show that the problem addressed in this study is not trivial and hence qualifies for a machine learning application. The remaining four are ML algorithms that are reportedly capable of handling the nonlinearity embedded in the data (Chen et al. 2019; Gillespie et al. 2021), especially ANN and SVM are well known for this (Xu et al. 2017). Other algorithms with nonlinear capabilities have been applied in the literature. However, these four are among the state-of-the-art having the best performance. Each of these algorithms is explained in the following subsections.

2.4.1. *Multivariate linear regression*

Multivariate linear regression (MLR) assumes a linear relationship between a set of input variables (measured soil quality parameters) and the target variable (soil quality indicator such as EC), often based on ordinary least squares framework (Xu et al. 2017, Anggraini et al. 2024). It estimates the coefficients of the linear eq. involving the input variables that best predict the values of the target variable but according to Aftab et al. (2022), MLR's prediction is "neither very accurate nor generalized". Though the estimated regression line is determined in a way that the square of the residuals is minimal as expressed in Alexopoulos (2010):

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \quad (2)$$

where Y is the target variable, β_0 is a constant value representing the intercept, $\beta_1 \ldots \beta_n$ are the regression coefficients, and $X_1 \ldots X_n$ are the input variables.

More details of the MLR algorithm can be found in Alexopoulos (2010). In a similar study, Tamal et al. (2021) applied one-way ANOVA

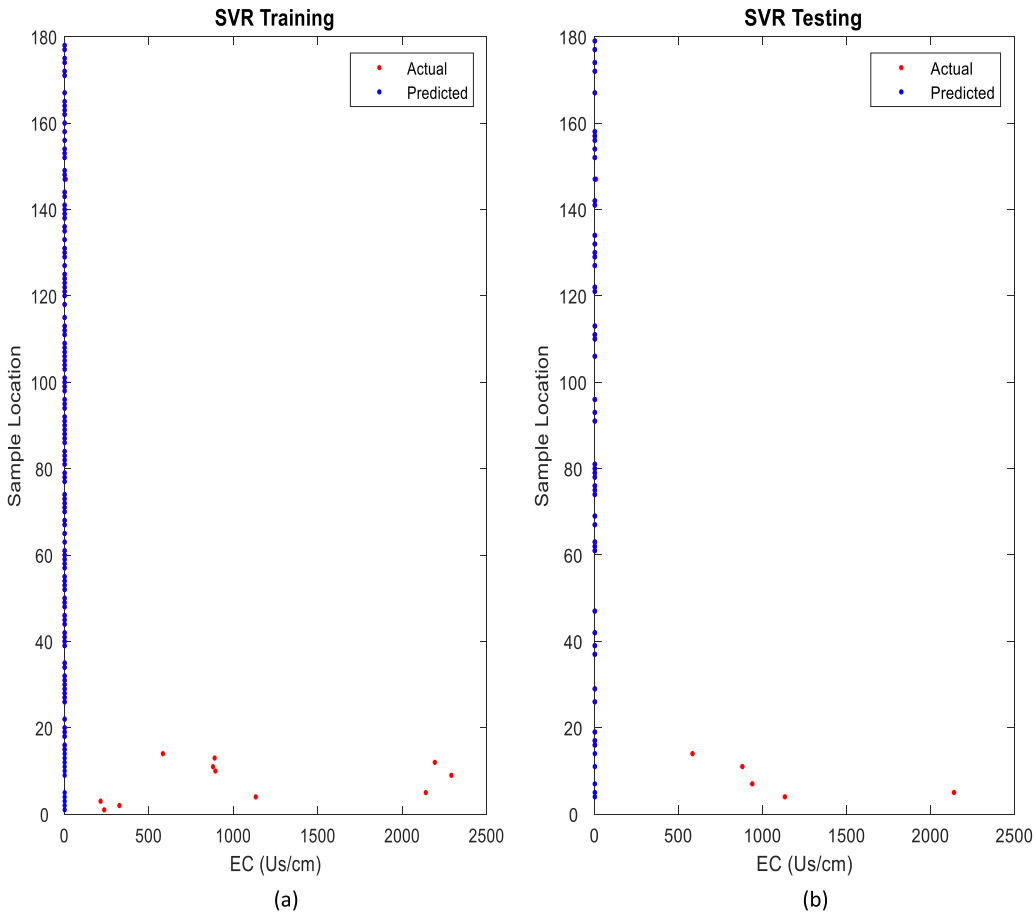

Fig. 5. Actual and Predicted (a) Training and (b) Testing Results of the SVR Model on Raw Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations. NB: The training and testing data were randomly selected using a stratified sampling approach where 70 % of the data was selected for training and 30 % reserved for blind test.

alongside SVM, k-nearest neighbor and ensemble bagged model.

2.4.2. *Artificial neural network*

ANN is a model inspired by an emulation of the biological systems. It is patterned after the interconnections between biological neurons and based on the human brain processing system (Sezer 2011). A neuron in ANN is a small computing unit that accepts inputs, processes them and produces an output. The ANN is composed of three layers: input, hidden and output. According to their relative importance, each input variable is multiplied by a weight assigned to it and gets "fired" to the hidden layer. The initial weights can be manually initialized or randomly assigned automatically. The latter is preferred to avoid human subjectivity. The hidden layer accepts the weighted inputs and processes them using a transfer function. The output layer is a discriminator. It accepts the processed signal from the hidden layer, checks how close they are to the actual output during the training process. If the error is greater than a certain threshold, the signals are propagated back to the input layer where the weights are adjusted to reduce the error (Priddy and Keller 2005). The cycle continues until the error goal is attained. According to Priddy and Keller (2005), a trained ANN model is represented as in eq. (3):

$$Y = \sum_{i=1}^{n} f(W_i X_i + b) \tag{3}$$

where Y is the output of the model, W_i are the weights assigned to each input variable X_i, and b is bias.

ANN has been used to identify key factors influencing variability of grain quality and corn yield based on soil properties such as soil conductivity (Papageorgiou et al. 2009). The reliability of ANN algorithms may be inferred from studies such as Alsharari et al. (2020), which obtained the most accurate prediction values from ANN model. Some ANN processes use neural network with random weights and it is believed to be highly effective (Song et al. 2020). More details of the ANN algorithm and its mathematical bases can be found in Priddy and Keller (2005), Beale et al. (2010); Sezer (2011); Zeng et al. (2017) and Bayatvarkeshi et al. (2020).

2.4.3. *Support vector regression*

SVR is a type of Support Vector Machine (SVM), a supervised machine-learning algorithm that is used for regression tasks (Liu et al. 2023). Instead of minimizing the observed training error, the main objective of SVR is to minimize the generalization error bound so as to achieve generalized performance. SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped through a nonlinear function. SVR handles nonlinear problems by pre-processing the training patterns while mapping them to some feature space. SVR is generalized as expressed in eq. (4) as simplified from the explanation of Awad and Khanna (2015):

$$Y = f(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) k(x_i, x) + b \tag{4}$$

where α_i^* and α_i are Lagrange multipliers whose coefficients are minimized by regularized risk function, $k(x_i, x)$ is the kernel function defined as a linear dot product of the nonlinear mapping of input variables x_i to a higher dimensional space, and b is an optional bias.

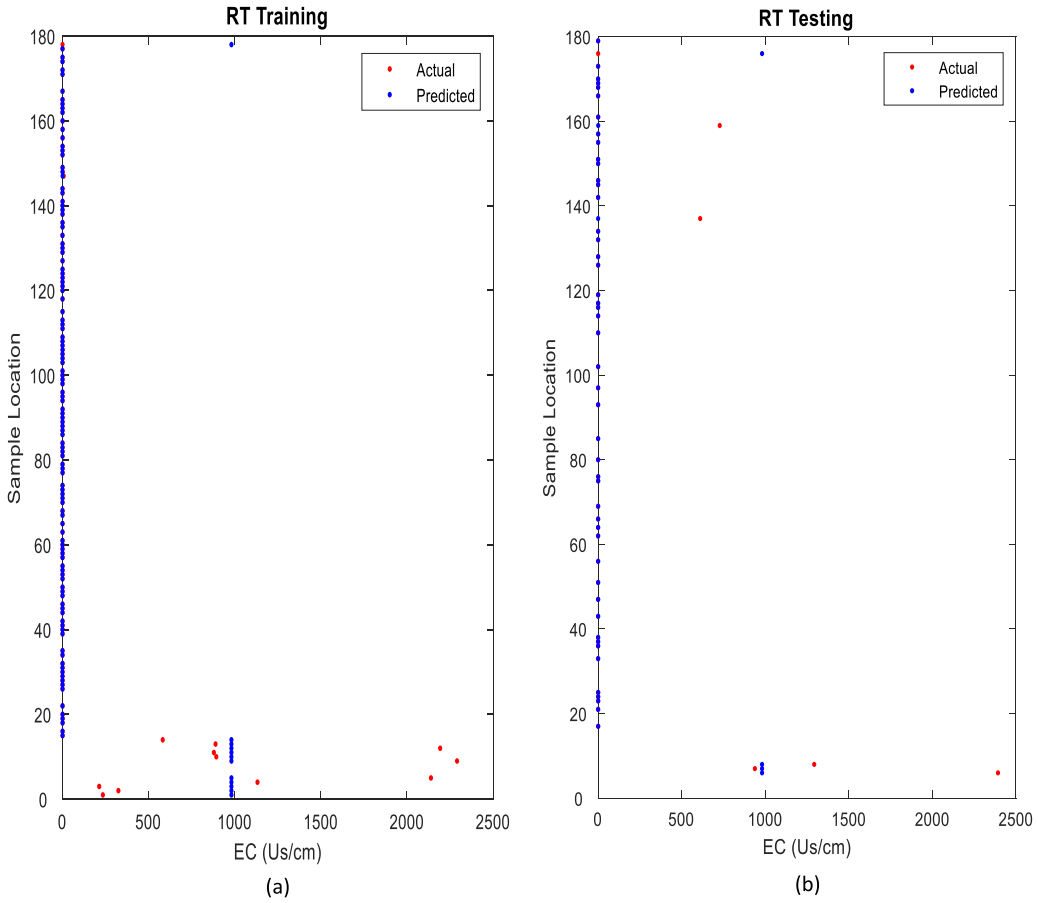

Fig. 6. Actual and Predicted (a) Training and (b) Testing Results of the RT Model on Raw Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations. NB: The training and testing data were randomly selected using a stratified sampling approach where 70 % of the data was selected for training and 30 % reserved for blind test.

More details about the SVR algorithm and its mathematical bases can be found in Basak et al. (2007); Awad and Khanna (2015); Zeng et al. (2017); Akinpelu et al. (2020) and Yin (2021).

2.4.4. *Regression tree / decision tree*

The regression tree (RT) algorithm is a supervised machine learning technique used for learning decision rules from features for both classification and regression tasks (Liberda et al. 2021). It uses a set of rules to predict future outcomes given input variables. It provides transparency for human interpretation by graphically representing the process of decision-making e.g., based on machine learning as demonstrated in Perboli and Arabnezhad (2021). It does this by creating a flowchart-like structure, and an if-else condition is applied at the nodes of the target attributes. The result of the rules is delivered at the leaf nodes (Breiman et al. 1984). The entire flow is from root to leaf. A regression tree works by dividing the predictor space into N distinct and non-overlapping regions. For any test or prediction observation, the mean of the training values that fall in the region is estimated. This is achieved by minimizing the entropy and maximizing the information gain (Yang et al. 2017). To minimize the prediction error, the generalized expression to predict a response variable is given in eq. (5) according to Yang et al. (2017).

$$Y = f(x) = \sum_{i=1}^{m} C_i . L_{x \in R_i} \tag{5}$$

where m is the number of regions into which the feature space is partitioned, C_i is the value chosen for the region R_i, L_x is the indicator function.

More details about the RT algorithm can be found in Breiman et al. (1984); Zeng et al. (2017); Yang et al. (2017) and Gillespie et al. (2021).

2.4.5. *Random forest*

The RF algorithm consists of a large number of individual decision trees that operate as an ensemble; and it is one of the popular tree-based classifiers (He and Fan 2021; Gillespie et al. 2021, Liu et al. 2023). Each individual tree in the RF algorithm produces a local output prediction and the average of the output predictions becomes the overall model prediction. The fundamental principle behind random forest is a simple but powerful one called the wisdom of crowds (Wang and Michoel 2019). The RF algorithm works by combining a large number of relatively uncorrelated tree models (Liu et al. 2023), called weak learners, operating as a committee and taking the average of the output of each individual committee member. The low or non-correlation, otherwise called diversity, between the individual models is the key to the working principle of random forest. Uncorrelated models have the capability to produce ensemble predictions that are more accurate than any of the individual predictions (Breiman 1996, 2001) and ensemble models are robust in scenarios with imbalance dataset (Perboli and Arabnezhad 2021). The reason for this effect is that the trees, with the diversity among them, protect each other from their individual errors. While some of the trees may be wrong, many other trees will be right, so as a group, the trees are able to move in the correct direction (Breiman 1996, 2001; Liu et al. 2023). According to Breiman (2001), the RF algorithm is generalized and expressed as:

$$Y = f(x) = \frac{1}{K} \sum_{k=1}^{K} h(x; \varnothing_k) \tag{6}$$

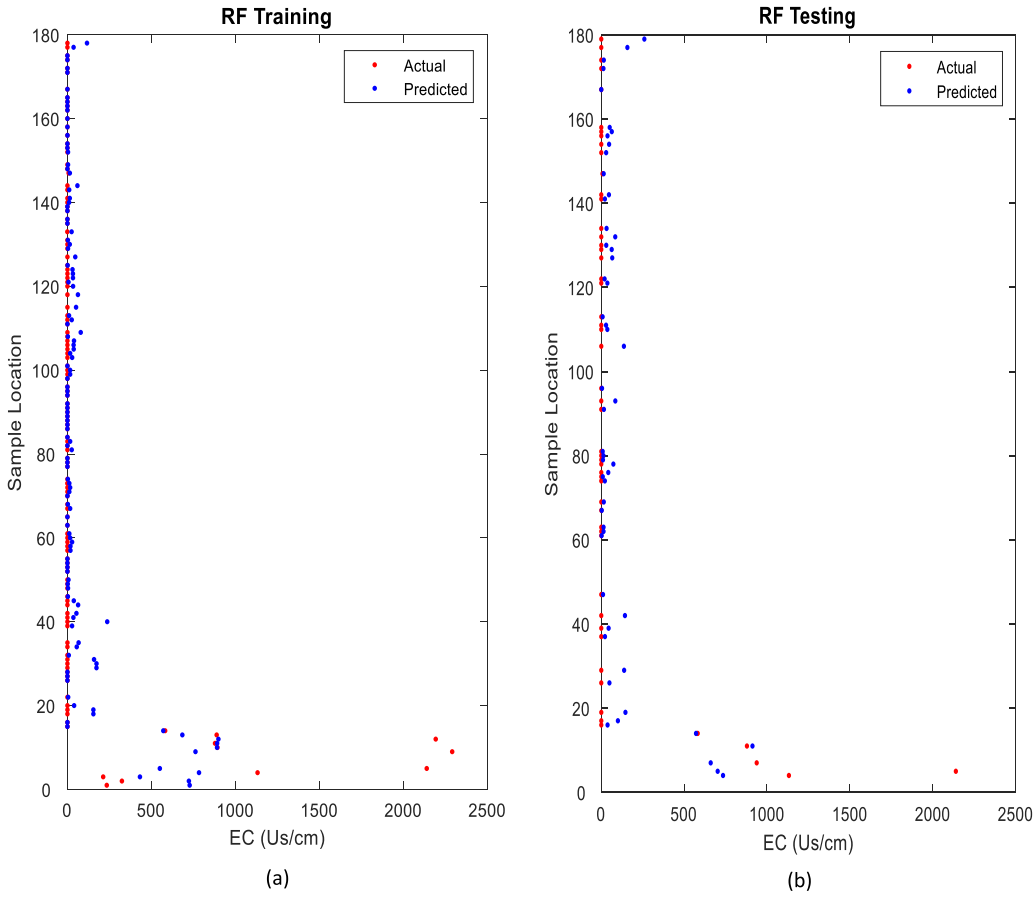

Fig. 7. Actual and Predicted (a) Training and (b) Testing Results of the RF Model on Raw Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations. NB: The training and testing data were randomly selected using a stratified sampling approach where 70 % of the data was selected for training and 30 % reserved for blind test.

where K is the number of weak learners, $h(x;\varnothing_k)$ is the collection of individual tree predictors, x represents the observed input covariate vector with its associated random vector, and the $\varnothing_k$ are independent and identically distributed random vectors.

More details about RF algorithm and its mathematical bases can be found in Breiman (1996, 2001); Shapire et al. (1998); Segal (2004); Wu et al. (2019); Brunswick et al. (2021) and He and Fan (2021).

2.5. Model performance evaluation criteria

This paper deployed the statistical performance and model evaluation criteria that are commonly used in the machine learning research community for regression. They are coefficient of determination (R-Square), p-value, and root mean square error (RMSE). R-Square indicates the proportion of the variance in the dependent variable that is predictable from the independent variable. It also ranges from 0 to 1. P-value is a statistical measure of the significance of the predictive power of a model. It is generally accepted that a p-value of ≤ 0.05 for a model means that the model is statistically significant (Greenland et al. 2016). This is interpreted to mean that the model has less than 1 in 20 chances of being wrong assuming the null hypothesis is true. The null hypothesis typically states that there is no relationship among the features used to build a model; and hence, not significant with respect to the contrary claim (Beaujean et al. 2010). The less the value, the better for a model as the sample data is said to provide enough evidence to reject the null hypothesis at the population level. The RMSE is the standard deviation of the prediction errors. It is a measure of how spread out the prediction errors are or how concentrated the predictions are around the line of best fit. For instance, Lai et al. (2021) detail its formulation and show how deep neural network application in solar radiation forecasting yielded a reduction in RMSE as opposed to its prevailing alternative.

3. Model design and implementation

3.1. Data Stratification

The datasets collected from 179 different locations all over the country (Fig. 1) consists of 25 individual measurements and were subjected to thorough and rigorous Quality Control (QC) at the desk in addition to those applied on the field. During the QC process, measurements with missing, unavailable, incomprehensible, or invalid entries were removed. These include Nitrates, Sulphate, Calcium, Phosphate, Nickel, Chromium, Mercury, Arsenic, Lead, Vanadium, Cadmium, Cyanide, Copper, and Zinc. After the QC process, only 10 measurements were left to be used as input variables while EC was used as the target. For a fair understanding of the dataset, Table 1 shows the basic statistical properties of the measurements.

From the statistical description presented in Table 1, it is apparent that all the measurements except pH and BTEX (to a lesser degree) have wide ranges that make them candidates for log-normalization. However, as will be further emphasized in the next section, this part of the modelling task is intended to confirm the performance of the proposed models on the raw data. The outcome of this exercise on whether log-normalization is required to improve the performance of machine learning models will add to future researchers' knowledge especially in this field.

The weak negative linear relationship between the measurements and EC as shown in Table 2, except for cl, could suggest that TPH, TAH,

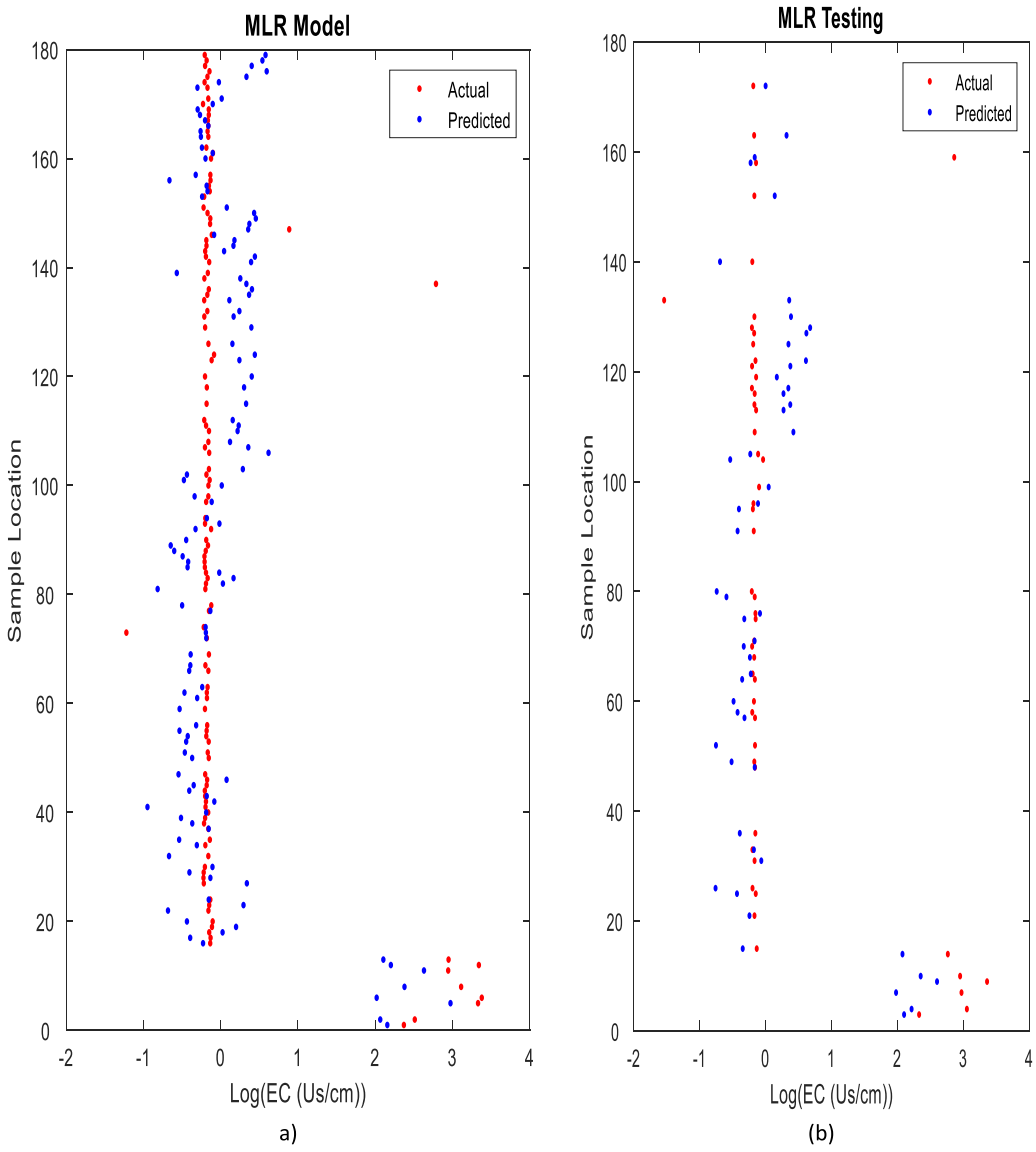

Fig. 8. Actual and Predicted (a) Training and (b) Testing Results of the MLR Model on Log-normalized Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations. NB: The training and testing data were randomly selected using a stratified sampling approach where 70 % of the data was selected for training & 30 % reserved for blind test.

PAH, and BTEX have little to do with EC. However, from previous studies (e.g. Mao et al. 2016), it is understood that lower hydrocarbon contents in fresh or biodegraded oil corresponds to higher EC. Any hidden nonlinear patterns and relationships in datasets are not always apparent to human. It is either challenging or out-rightly impossible to discover such hidden patterns in the empirical approach mostly being employed in the industry. Hence, the application of machine learning to explore and leverage these unseen nonlinear relationships as may have been revealed in section 3.2.2.

It is important to emphasize that we did not "select" the predictors. Rather, the study maximised the capability of ML algorithms, given their sophisticated mathematical and sound statistical bases, to handle multivariate and multidimensional data. They also have the capability to discover the nonlinear patterns that are hidden and beyond the limited human capability. These smart algorithms were used not only to handle all the data available but also use their capability to automatically select the most optimal predictors for the target. Doing this tends to remove human bias and subjectivity from the feature selection process.

Following the standard machine learning procedure, 70 % of the dataset was randomly selected for training and the remaining 30 % was reserved for blind test. Various stratification percentages have been used in the literature (Zeng et al. 2017). Our choice of this 70:30 ratio is based on its recommendation in various published studies such as Anifowose and Abdulraheem (2011). It should be noted that this is the best achievable method of data stratification in the face of the data sparsity problem encountered in this study. The random selection is based on a stratified sampling scheme. This scheme ensures a random selection of the training and testing subsets such that each sample has an equal chance of being selected, which limits bias. Using this stratification scheme, from the total of 179 samples, 126 were selected for training while 53 were used for blind test to evaluate the performance of the models. On top of the stratified sampling approach, the k-fold cross-validation was applied on the training subset to avoid overfitting.

3.2. *Model design and optimization*

The MLR model with optimized coefficients and it was implemented as a benchmark to demonstrate that this problem is not a trivial one that could be modelled by a simple linear correlation method. For the ANN model, the optimal configuration consists of a training function that uses

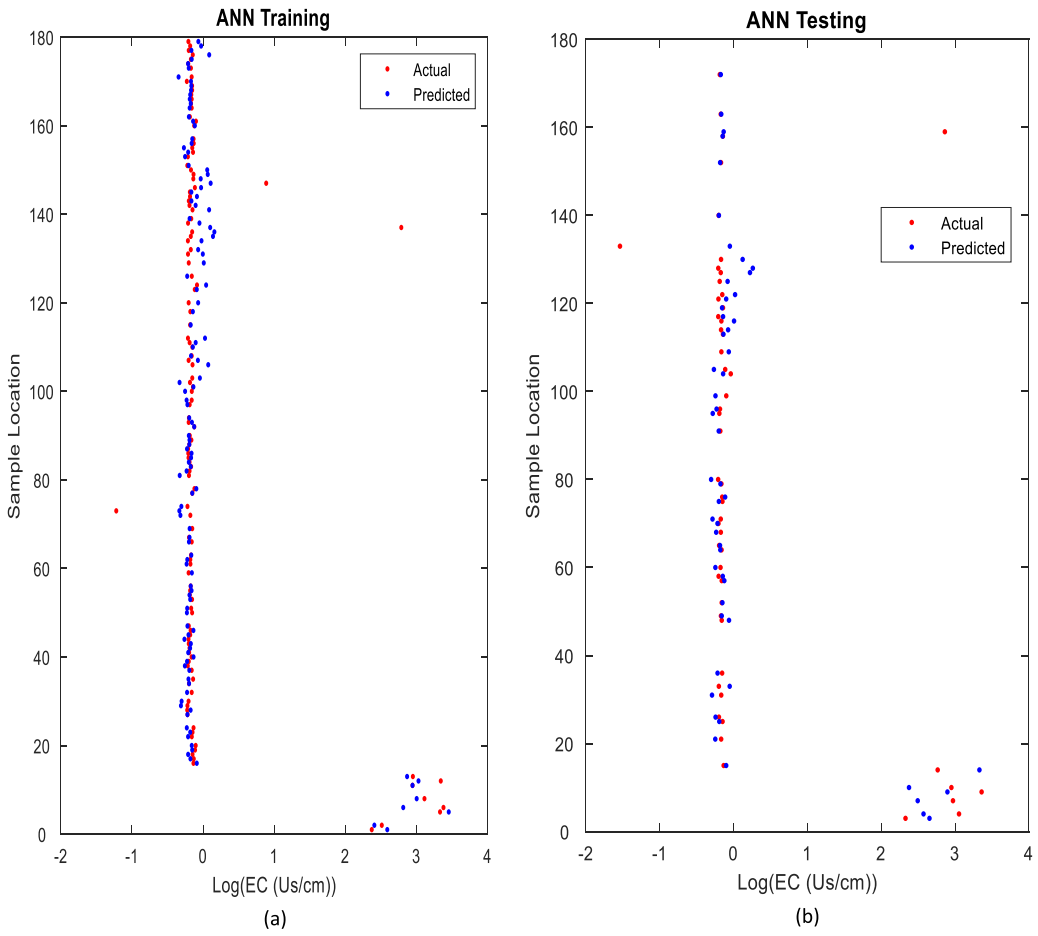

Fig. 9. Actual and Predicted (a) Training and (b) Testing Results of the ANN Model on Log-normalized Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations. NB: The training and testing data were randomly selected using a stratified sampling approach where 70 % of the data was selected for training & 30 % reserved for blind test.

the Levenberg-Marquardt backpropagation (trainlm), one hidden layer, and 10 neurons in the hidden layer. This emerged after different training functions, number of layers, and number of neurons in the hidden layer were tested and evaluated. To reduce the effect of local minimum on the results typical of ANN as widely reported in the literature (de Weijer 1993; Nag et al. 2005; Saad and Wunsch 2007; Huang et al. 2018), 50 runs were conducted and the average of the runs was taken as the overall prediction. This is an ensemble approach and the number "50" is chosen arbitrarily but increasing the number could be better (Kocsis et al. 2013; Htike 2016).

For the SVM model, the optimal configuration consists of a Gaussian kernel function while the regularization (i.e. boxconstraint) and epsilon parameters were determined automatically by taking the interquartile range (IQR) of the target values. The Gaussian kernel function emerged after several other functions such as polynomial, tangential, and hyperbolic were tested and evaluated. For this problem, the IQR is the difference between the 75th and 25th percentiles of EC data. The IQR is a robust estimate of the spread of the data, since changes in the upper and lower 25 % of the data do not usually affect the spread (Chau et al. 2005; Sarma 2018; Jones 2019). For the RT model, an optimized, cross-validated and pruned tree with surrogate set on and a minimum leaf size of 12 were used. This emerged after several minimum leaf sizes were tested and evaluated. Cross-validation, surrogacy, and pruning are well known methods of achieving the best training performance of regression trees.

For the RF model, a bagged ensemble of regression trees (e.g., a type of regression ensemble grown by resampling, which combines predictions from various decision trees) with a minimum leaf size of 8 and 30 learning cycles were used. This optimum level was achieved after several trees, minimum leaf sizes, and learning cycles were tested and evaluated. The modelling process is conducted in two parts. The initial step uses the entire dataset without normalization while the second part uses the log-normalized versions of those measurements identified to have high variability, including EC, based on the results presented in Table 1.

3.2.1. *Results of modelling with raw input measurements*

For the initial run of the models, all the measurements were used in their raw conditions and no pre-processing or transformation was applied. This is an exploratory research decision that aims to compare the results of using the raw measurements to the normalized version. Since some ML methods and certain data conditions are not affected by normalization, the intension was to explore this in the context of this application. The results obtained are summarized in Table 3 and the comparative plots are presented in Fig.s 3 to 7.

As shown in Table 3, the RF method showed the highest resilience in the ability to handle the raw measurements with wide variability without log-normalization, just as Zeng et al. (2017) found RF the best model against RT, SVM and ANN. Both MLR and SVR are underfitted. This ML phenomenon is indicative of the situation when both the training and testing performances are low. This behaviour is typical of MLR that is only capable of handling linear relationships (Aftab et al. 2022). For SVR, it could be that the optimized parameters were not sufficient to approximate the input function to match the target function unlike its hybrid application using manual search and genetic algorithm in Akinpelu et al. (2020). This equally applied to the other methods

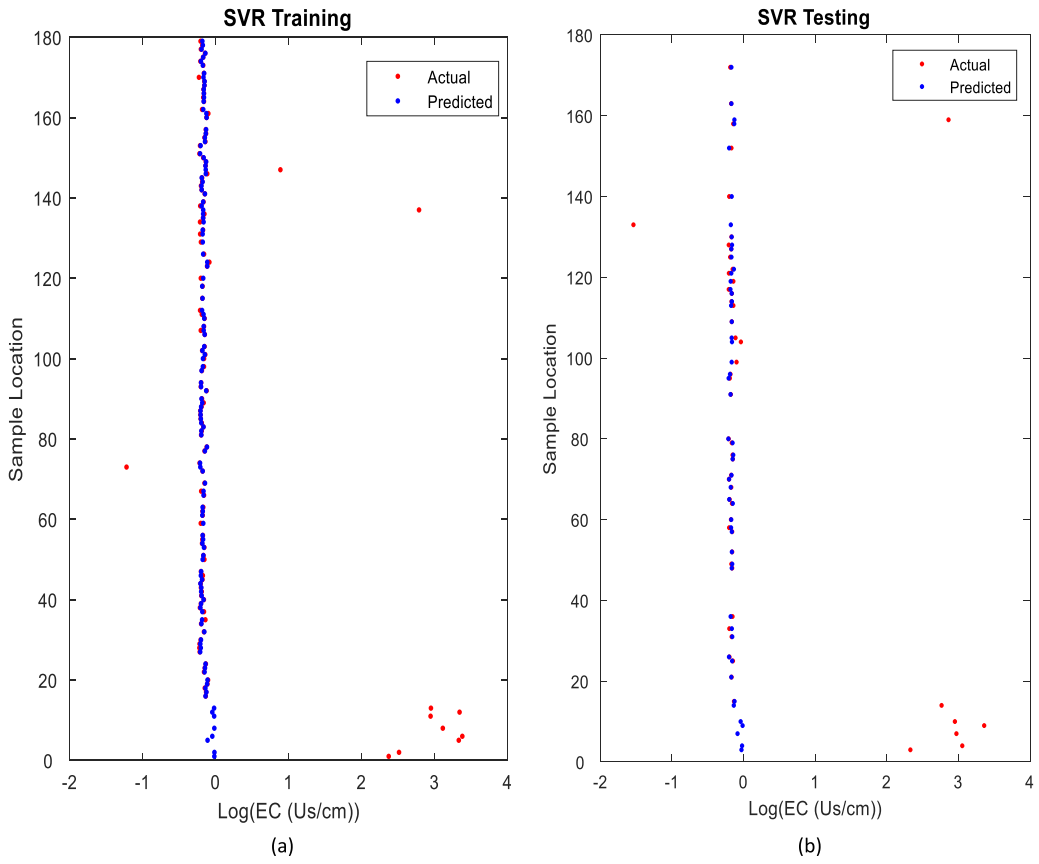

Fig. 10. Actual and Predicted (a) Training and (b) Testing Results of the SVR Model on Log-normalized Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations. NB: The training and testing data were randomly selected using a stratified sampling approach where 70 % of the data was selected for training & 30 % reserved for blind test.

though to a lesser degree, leading to higher disparities between the actual and predicted EC values. Yin (2021) found that in comparison to other methods, SVM minimises failures by improving performance against bribery and comprise in the regulatory frameworks governing offshore oil and gas exploitation. Table 3 shows R-square values for SVR training and testing models are closer to zero unlike in Aftab et al. (2022).

Following the poor training and generalization performances of the MLR and SVR models, their p-values, especially for the testing, indicated that they would not make good models. This confirms the utility of p-value as an indication of model goodness-of-fit. The real effect of using the raw measurements despite their wide variability is demonstrated by the RMSE values. It would be noted that while R-Square measures the degree of collinearity between the actual and predicted EC, the p-value gives indication of the goodness-of-fit for the models, and the RMSE shows the residual between the actual and predicted EC. Each of these evaluation criteria could not have been used in isolation (Aftab et al. 2022). This shows how numerically far apart the predicted EC values are to the actual. It was therefore concluded that using the raw measurements with such wide variability in their values is not advisable. This conclusion is confirmed by Fig.s 3 to 7 showing a high degree of disparity between the actual and predicted EC values. To guide readers' interpretation of Fig.s 1 to 7, it is important to note that the red dots are the actual EC measurements while the blue are the predicted values (this aligns with the legend). The closer the red and blue dots on the same line, the better the prediction result. Similarly, the farther the dots, the higher the prediction error.

Confirming the results shown in Table 3, it is observed in Fig.s 3 to 7 representing the comparative plots of the actual and the predicted EC values for MLR, ANN, SVR, RT, and RF respectively that majority of the actual EC values are close to zero. This affected the training capabilities of the MLR (Fig. 3) and SVR (Fig. 5) models. The MLR model could not handle the nonlinearity unlike RF algorithm (Anggraini et al. 2024) while the SVR model became too attached to the low EC values. This made it challenging for both models to train and, to a worse degree, generalize effectively on the low EC values. With the high variabilities in the EC values, the disparities led to the high RMSEs. The other models, ANN (Fig. 4), RT (Fig. 6), and RF (Fig. 7), showed good agreement between the actual and predicted EC values in most of the sample locations but with high variations especially in the sample locations below 20. More details about these results are discussed in Section 4. To reduce the adverse effect of the high variability in the EC values, the appropriate measurements were log-normalized. The next section presents the results of modelling using the log-normalized input measurements.

3.2.2. *Results of modelling with log-normalized input measurements*

To improve the performance of the models and using the statistical description of the dataset presented in Table 1 as a guide, a second run was conducted but with the measurements having a high degree of variability indicated by the range up to three orders of magnitude log-normalized as recommended in Hou et al. (2020). The measurements that are affected by this decision are: Redox, Na, TPH, TAH, PAH, Fe, and EC. To further achieve the log normalization, there was the need to remove PAH and BTEX since their log values are either undefined or infinite (-inf). With this process, only eight measurements were left for modelling. Using the log-normalized measurements, the results obtained are summarized in Table 4 and the comparative plots presented in Fig.s 8 to 12.

The results presented in Table 4 show that the log-normalization process greatly improved the performance of the models. The MLR

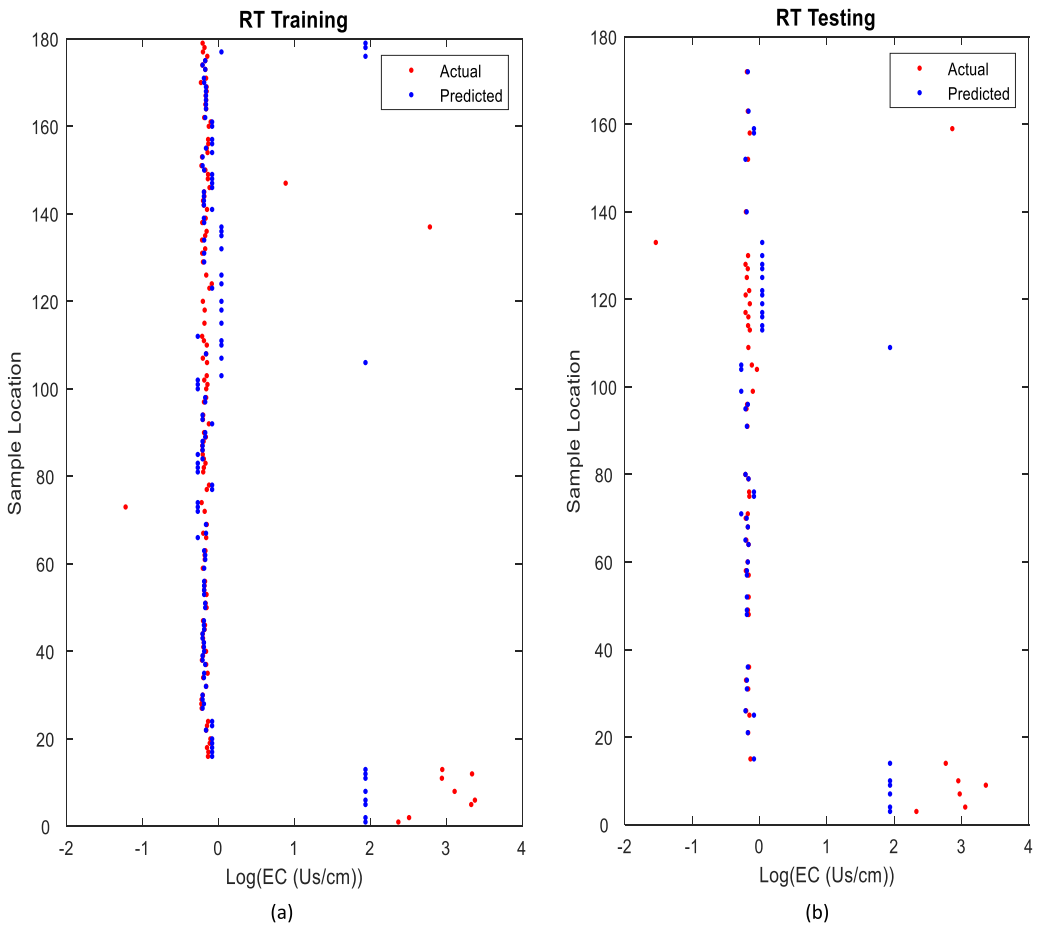

Fig. 11. Actual and Predicted (a) Training and (b) Testing Results of the RT Model on Log-normalized Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations. NB: The training and testing data were randomly selected using a stratified sampling approach where 70 % of the data was selected for training & 30 % reserved for blind test.

and SVR models that were under fitted with the raw data (Table 3, Fig.s 3 and 5) are now not only sufficiently trained but also able to generalize on the blind data subset. Though ANN and SVR are extremely complex, the latter can be improved by inclusive ensemble model (Lin and Billa 2021) while the effect of the log-normalization on reducing the high variability in the EC values has led to the reduction in the disparities between the actual and the predicted values (Fig.s 8 to 12). This reduced the RMSE, just as RMSE displayed the best performance in Lai et al. (2021). Matsui et al. (2022) suggest there could be future possibility to divert trained models to other languages e.g., as part of deep-learning NLP technique.

Confirming the improved performance of the models (Table 4 and Fig.s 8 to 12) representing MLR, ANN, SVR, RT, and RF, it shows that there is increased correlation unlike in Table 2 and lower error between the actual and predicted EC values. For the MLR model (Fig. 8), its incapability to handle the nonlinearity of the relationship between EC and the input measurements became evident as indicated by the disagreements in most locations between the actual and the predicted EC values. For the ANN model (Fig. 9), there is good agreement between the actual and the predicted EC values in most locations except in locations between 130 and 160 for both training and testing data. For the SVR model (Fig. 10), there are good agreements between the actual and the predicted EC values not in most locations but very wide variations not only in the locations between 130 and 160, but also in those below 20. The RT model (Fig. 11), despite good agreement between the actual and the predicted EC values in most locations, had a lot of variations in more locations between 70 and 160 in addition to those below 20. The RF model (Fig. 12) showed the most agreement between the actual and the predicted EC values in most locations including those below 20. This contributed to its emerging as the most acceptable of all the models. More details about these results are also discussed in Section 4.

Summarizing the comparative results, Fig.s 13 to 15 show the comparisons of the R-Squares, p-Values, and RMSE for all models with respect to the raw and log-normalized datasets. A more comprehensive analysis of the summarized results is presented in Section 4.

4. Further Discussion and Analysis of Results

The results presented in Table 3 show that despite the wide variability that exists in some of the measurements, the machine learning models (Fig. 4 to 7) perform better than the MLR model (Fig. 3), as demonstrated in Lin and Billa (2021) and Anggraini et al. (2024). This is an indication that the nature of soil impurities as observed through the EC patterns cannot be explained by a linear relationship just as Mao et al. (2016) found that an increase in conductivity corresponds to a decrease in total PH. Among the machine learning models (Fig. 4 to 7), the RF model (Fig. 7) that is based on ensemble learning methodology demonstrated the best performance (Wu et al. 2019) as indicated by the highest R-Square and the lowest RMSE values for the testing evaluation. This agrees with similar reports in the literature (e.g., Segal 2004; Zeng et al. 2017; Wang Y. et al. 2019; Chen et al. 2019; Khan et al. 2019; Hou et al. 2020; Aftab et al. 2022). The next in the performance rating is the RT model (Fig. 6). This shows that the single tree model is good in itself but could perform better when combined in an ensemble fashion. This also agrees with the literature on the good performance of decision tree models for regression or classification (Singh et al. 2013; Pandey and

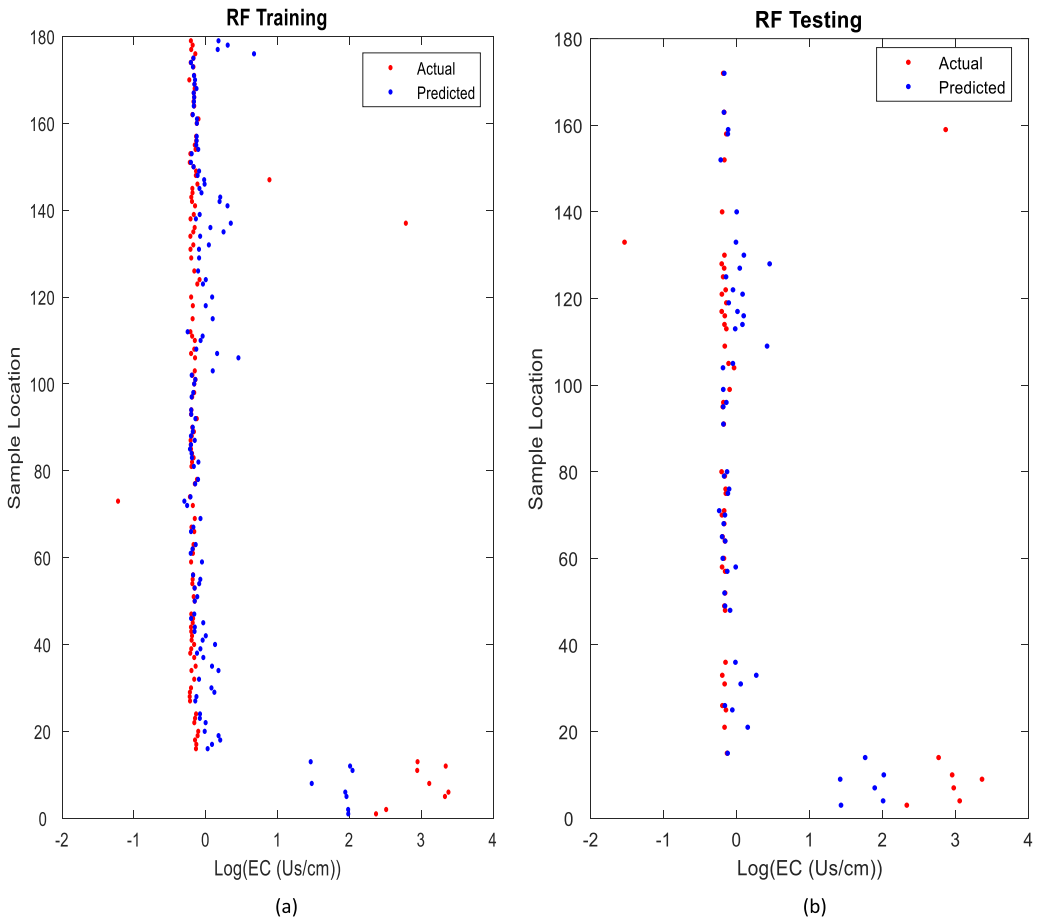

Fig. 12. Actual and Predicted (a) Training and (b) Testing Results of the RF Model on Log-normalized Measurements of heavy metals, petroleum hydrocarbons and EC from sampled oil facility locations. NB: The training and testing data were randomly selected using a stratified sampling approach where 70 % of the data was selected for training & 30 % reserved for blind test.

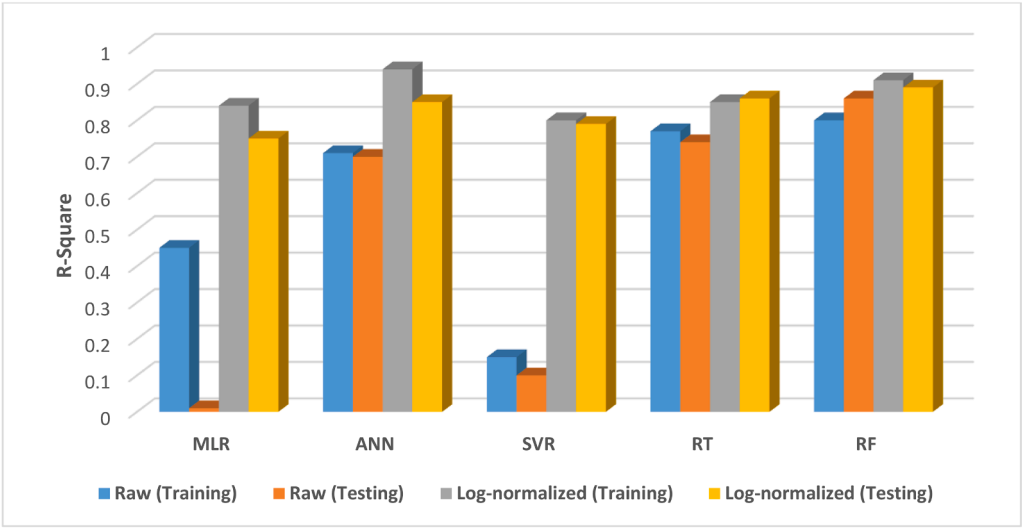

Fig. 13. R-square comparison of all models for raw and log-normalized datasets

Sharma 2013).

The ANN model (Fig. 4) also proved to be good despite its challenge with the local optimum, as Chen et al. (2019) found ANN model to have performed moderately worse. We attribute the good performance to our approach of running 50 cycles of the model, in an ensemble fashion, and taking the average as the overall prediction values. This approach proved to be capable of neutralizing the effect of the local optimum. This further confirms the efficacy of the ensemble learning approach to modelling, which makes them attractive as the researcher is not burdened by the choices of final exposure assignment model (Chen et al. 2019). ANN is known to be significantly efficient in predicting soil pollution parameters (Mojid et al. 2019) but not with such a complex

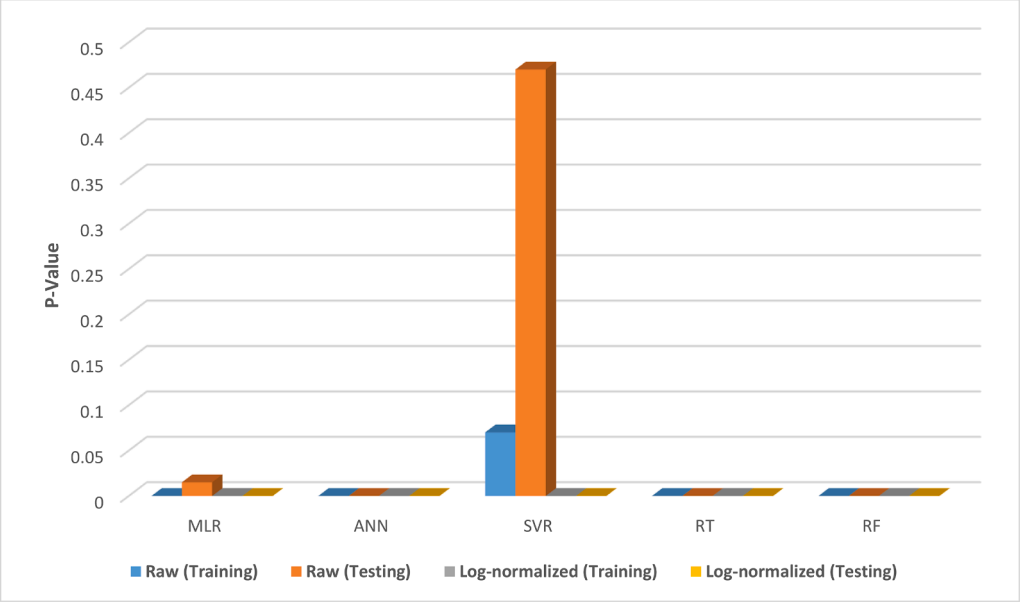

Fig. 14. P-value comparison of all models for raw and log-normalized datasets

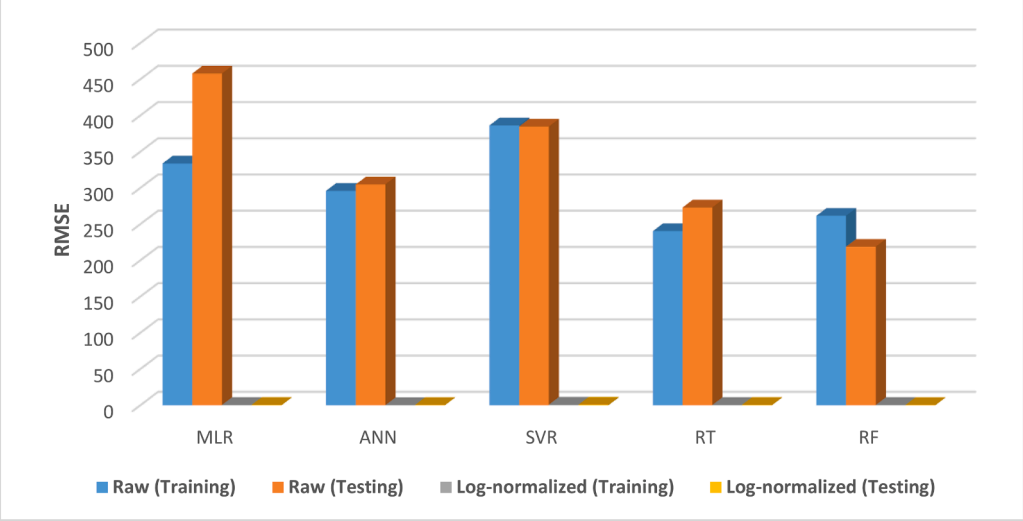

Fig. 15. RMSE comparison of all models for raw and log-normalized datasets

mix of pollutants as demonstrated in this paper. The SVR model (Fig. 5) showed poor performance and it was the second worst performing in a similar study by Zeng et al. (2017). This is contrary to its excellent performance reported in the literature (Zhu et al. 2010; Anifowose et al. 2011; Ewenla et al. 2013). The suspicion is that the poor SVR performance herein is due to sparsity of some samples in the dataset; and Hou et al. (2020) suggest that SVM is one of those machine learning models sensitive to the scale of data. Despite that, data sparsity is a general problem in the machine learning community and this problem has been reported with direct reference to Support Vector Machines in particular (Li et al. 2015). Our future work will further explore new approaches to handle data sparsity in machine learning and especially for SVR.

The effect of the variability in the data and especially for EC is evident in the results presented in Fig.s 3 to 7. Since the range of the target values are large, a little deviation in the prediction values resulted in significant difference, hence large RMSE values. This is the major reason for the high RMSE values presented in Table 3 even for the good performing models. Despite the good performance of the ANN (Fig. 4), RT (Fig. 6), and RF (Fig. 7) models, the high RMSE associated with them would not make them good models to efficiently generalize and be trusted on new measurements for future predictions (Huo et al. 2019). This is supported by the high p-values exhibited by the models as will be discussed subsequently. The occurrence of high RMSE in the initial models necessitated the second run of the models with the relevant measurements, including EC, log-normalized. This process helped to reduce the high variability that existed in the measurements, and it significantly improved the results of the models including the SVR (Fig. 5).

The log-normalization of relevant measurements significantly improved the predictive power of the models as evident in their higher R-Square values and much lower RMSE values than those exhibited for the raw dataset. Table 4 shows that the RT model (Fig. 11) maintained its best performance with the highest R-Square and the lowest RMSE values. The performance metric of the RF model (Fig. 12) is closely followed by those of RT (Fig. 11) and ANN (Fig. 9) models. Focusing on the SVR model (Fig. 10) revealed that its performance is significantly

improved by the log-normalization process. Even though the performance of the MLR model improved significantly, its lowest R-Square and RMSE values proved that the nonlinearity in the relationship between the EC and other soil quality measurements is better analysed by machine learning models rather than such a linear model as MLR. As a direct consequence of the log-normalization process, the p-values of the models improved as the values are now much lower than previously. The effect of the log-normalization process is also seen in the prediction plots of the models as shown in Fig.s 8 to 12. Now the actual and predicted values are closer together and a little deviation in the prediction does not result in high prediction error anymore. Perhaps, the justification of the claim that ML application "ignores soil science knowledge" including the possibility of misleading and wrong results (Rossiter 2018; Padarian et al. 2020) can be found in the MLR performances compared to the other models in both the raw and log-normalised experiments, since the "hidden" nonlinear patterns in the datasets are generally discernible by ML only, not human.

Overall, Fig. 13 shows the significant improvement in the performance of the models especially the MLR and SVR as a direct consequence of the log-normalization process. The other models also increased in their performance as demonstrated in their increased R-Square values despite their high R-Square values before the log-normalization process. Fig. 14 strengthens the demonstration of performance improvement of the models with special focus on the MLR and SVR as their p-values reduced considerably. Fig. 15 agrees with the previous performance results by showing the significant and drastic reduction in the prediction errors with up to three orders of magnitude. These results altogether attest to the positive impact of log-normalization in data analysis and machine learning modelling.

Although Mao et al. (2016) found that lower hydrocarbon content in fresh or biodegraded oil corresponds to higher EC in lab experiment, the hitherto hidden nonlinear relationships between EC and petroleum hydrocarbons are shown in Fig.s 4 to 12. Their work further shows that, as average, EC ratio increases 69 days after the start of the laboratory experiment, it slowly decreased after 187 days. As with other studies, TPH measurement shows the presence (or otherwise) of petroleum hydrocarbons in any sampled medium; hence it is an anthropogenic impurity useful for examining soil EC variations. BTEX is a specific contaminant known to be a component of TPH unlike PAH and TAH. The laboratory quantification of PAH is far more difficult than TPH as the latter is easily determined through experiments (Akinpelu et al. 2020).

5. Conclusions and future work

For the first time, this study demonstrates the efficacy of machine learning techniques in teasing out the nonlinear patterns in soil data containing complex mixture of heavy metals, petroleum hydrocarbons and physicochemical parameters in a data-sparse tropical region. This is significant for the hypothesis that ML application "ignores soil science knowledge", and this hypothesis was largely based on soil data from the temperate region as opposed to the tropical region data employed in our study. Therefore, the results from our study appear to support this hypothesis as ML approaches (ANN, RF, SVR, RT) needed no human intervention in the processing of the 10 independent variables and their statistical relationships with EC. The computer power with the choice of covariates built the models without pedological knowledge or an understanding of soil characterisation. In addition, this study presents novel interpretations, and machine-learning analyses, of soil experimental data from the data sparse region of Africa's largest oil producer, Nigeria. Also, robust models with excellent predictive capabilities were built to reduce uncertainties and increase the accuracy of environmental impact prediction as would be expected of EIAs. The insights obtained from the study results are:

i. The estimation of EC from soil quality measurements is nonlinear. This is the reason the MLR model performed less than the machine learning models.
ii. Despite the wide variability in the distribution of the raw dataset, some machine learning models (ANN, RT, and RF) can handle the challenges of noisy and nonlinearity of the data.
iii. Log-normalization helped to improve the predictive capability of all models by removing the effects of statistical variability in the dataset.
iv. Machine learning has potential application in environmental impact prediction, and by extension, can engender better management decisions toward the sustainable development of natural resources through the Environmental Impact Assessment (EIA) process. This has a significant implication for global EIA practitioners and researchers as the US-based International Association for Impact Assessment – IAIA (with thousands of members across 120 countries) continues to explore opportunities to deploy AI-ML techniques to improve impact prediction processes.
v. Although ML application possibly "ignores soil science knowledge" (Rossiter 2018) while its results could be "misleading and wrong" (Padarian et al. 2020), our findings from a data-sparse (rarely studied) tropical region contribute to the growing debate.

The implication of this study is that data sparsity is, increasingly, no longer an excuse for the absence of quantitative impact prediction in EIA processes as this has been a major problem reported in the literature (e.g., see Ogunba 2004; Anifowose et al. 2016). Though study limitations include the use of data collected by third parties with some data points missing, unavailable, or invalid (see Section 3.1), future effort will be directed towards exploring various algorithms to handle data sparsity to further increase the predictive power of the models. In our future work, the expectation is that combining the log-normalization process with data sparsity reduction approaches will help to unravel the mystery behind the poor performance of the SVR model during the first run using the raw dataset. Our future work will also explore various nonlinear feature selection algorithms that will recommend a subset of the input variables to yield significant improvement in the predictive power of the models.

Thus far, this study and its results demonstrate that the emergence of data science (i.e., the act of extracting meaningful insights from data using AI, ML) alongside interdisciplinarity (e.g., fusion of the fields of impact assessment, computer science, environmental management, and environmental science) can encourage innovative new approaches to solve complex sustainability problems (e.g., see Pennington 2020).

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statements

The data that support the findings of this study are available from Petroleum Products Marketing Company (PPMC) / Bureau of Public Enterprises (BPE) but restrictions may apply to the availability of these data, which were used under agreement for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of PPMC/BPE.

CRediT authorship contribution statement

Babatunde Anifowose: Writing – review & editing, Writing – original draft, Visualization, Investigation, Formal analysis, Data curation, Conceptualization, Methodology. **Fatai Anifowose:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis.

Declaration of competing interest

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

The lead author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for [Frontiers in Environmental Engineering] and was not involved in the editorial review or the decision to publish this article.

Data availability

Data will be made available on request.

Acknowledgement

The authors are grateful for the suggestions and constructive feedback from six anonymous peer-reviewers and the journal's Editor-in-Chief. We acknowledge the teams that undertook the field work, the environmental audit and the collaborating agencies including late Dr (Mrs) M.T. Odubela of the Environmental Assessment department in the Federal Ministry of Environment, Abuja. The Federal Government of Nigeria through the National Council on Privatization (NCP) and the Bureau of Public Enterprises (BPE) commissioned the environmental audit of the PPMC facilities and was co-funded by the World Bank.

References

Abdel Aal, G.Z., Atekwana, E.A, 2014. Spectral induced polarization (SIP) response of biodegraded oil in porous media. Geophys. J. Int. 196 (2), 804–817.

Aftab, R.A., Zaidi, S., Danish, M., Ansari, K.B., Danish, M., 2022. Novel Machine Learning (ML) models for predicting the performance of multi-metal binding green adsorbent for the removal of Cd (II), Cu (II), Pb (II) and Zn (II) ions. Environ. Adv. 9. https://doi.org/10.1016/j.envadv.2022.100256.

Anggraini, T.S., Irie a, H., Sakti, A.D., Wikantika, k., 2024. Machine learning-based global air quality index development using remote sensing and ground-based stations. Environ. Adv. 15, 100456.

Ai, D., Yang, J., 2019. A machine learning approach for cost prediction analysis in environmental governance engineering. Neural Comput. Appl. 31, 8195–8203.

Akinpelu, A.A., Ali, M.E., Owolabi, T.O., Johan, M.R., Saidur, R., Olatunji, S.O., Chowdbury, Z., 2020. A support vector regression model for the prediction of total polyaromatic hydrocarbons in soil: an artificial intelligent system for mapping environmental pollution. Neural Comput. Appl. https://doi.org/10.1007/s00521-020-04845-3.

Alexopoulos, E.C., 2010. Introduction to multivariate regression analysis. Hippokratia 14 (S1), 23–28.

Alsharari, B., Olenko, A., Abuel-Naga, H., 2020. Modeling of electrical resistivity of soil based on geotechnical properties. Expert. Syst. Appl. 141, 112966.

Anifowose, F., Ewenla, A., Eludiora, S., 2011. Prediction of oil and gas reservoir properties using support vector machines. In: Manuscript ID# IPTC-14514-PP: 2011 International Petroleum Technology Conference (IPTC). Bangkok, Thailand, pp. 15–17.

Anifowose, F., Abdulraheem, A., 2011. Fuzzy logic-driven and SVM-driven hybrid computational intelligence models applied to oil and gas reservoir characterization. J. Nat. Gas. Sci. Eng. 3 (3), 505–517.

Anifowose, B.A., Lawler, D.M., van der Horst, D., Chapman, L., 2012. Attacks on oil transport pipelines in Nigeria: a quantitative exploration and possible explanation of observed patterns. Appl. Geography 32, 636–651.

Anifowose, B.A., Lawler, D.M., van der Horst, D., Chapman, L., 2014. Evaluating interdiction of oil pipelines at river crossings using Environmental Impact Assessments. AREA 46 (1), 4–17.

Anifowose, B., Lawler, D.M., van der Horst, D., Chapman, L., 2016. A systematic quality assessment of Environmental Impact Statements in the oil and gas industry. Sci. Total Environ. 572, 570–585.

Anifowose, F., Labadin, J., Abdulraheem, A., 2017. Ensemble machine learning: An untapped modeling paradigm for petroleum reservoir characterization. J. Petrol. Sci. Eng. 151, 480–487.

Anifowose, B., Odubela, M., 2018. Oil facility operations - a multivariate analysis of water pollution parameters. J. Clean. Prod. 187, 180–189.

Awad, M., Khanna, R., 2015. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Apress Open, p. 263.

Basak, D., Pal, S., Patranabis, D.C., 2007. Support vector regression. Neural Inf. Process. Lett. Rev. 11 (10), 203–224.

Bayatvarkeshi, M., Mohammadi, K., Kisi, O., Fasihi, R., 2020. A new wavelet conjunction approach for estimation of relative humidity: wavelet principal component analysis combined with ANN. Neural Comput. Appl. 32, 4989–5000.

Beale, M.H., Hagan, M.T., Demuth, H.B., 2010. Neural Network Toolbox User's Guide. The Mathworks Inc., Natick, MA, USA.

Bieganowski, A., Józefaciuk, G., Bandura, L., Guz, Ł., Łagód, G., Franus, W., 2018. Evaluation of hydrocarbon soil pollution using E-Nose. Sensors 18, 2463.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees, 1st Edition. Chapman and Hall/CRC, p. 368.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Brunswick, P., Cuthbertson, D., Yan, J., Chua, C.C., Duchesne, I., Isabel, N., Evans, P.D., Gasson, P., Kite, G., Bruno, J., van Aggelen, G., Shang, D., 2021. A practical study of CITES wood species identification by untargeted DART/QTOF, GC/QTOF and LC/QTOF together with machine learning processes and statistical analysis. Environ. Adv. 5 https://doi.org/10.1016/j.envadv.2021.100089.

Chau, T., Young, S., Redekop, S., 2005. Managing variability in the summary and comparison of gait data. J Neuroengineering Rehabil. 2 (22), 1–20. https://doi.org/10.1186/1743-0003-2-22.

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

Chou, J.S., Truong, D.N., Le, T.L., Truong, T.T.H, 2021. Bio-inspired optimization of weighted-feature machine learning for strength property prediction of fiber-reinforced soil. Expert. Syst. Appl. 180, 115042.

Cunat, P-J. 2002. Corrosion Resistance of Stainless Steels in Soils and in Concrete. Paper presented at the Plenary Days of the Committee on the Study of Pipe Corrosion and Protection - Stainless Steels in Soils and in Concrete. Available at: http://www.worldstainless.org/Files/issf/non-image-files/PDF/Euro_Inox/CorrResist_SoilsConcrete_EN.pdf. Accessed on: 22 August 2019.

da Silveira, V.A., Veloso, G.V., de Paula, H.B., dos Santos, A.R., Schaefer, C.E.G.R., Fernandes-Filho, E.I., Francelino, M.R., 2022. Modeling and mapping of Inselberg habitats for environmental conservation in the Atlantic Forest and Caatinga domains. Brazil. Environ. Adv. 8 https://doi.org/10.1016/j.envadv.2022.100209.

de Weijer, A.P, Lucasius, C.B., Buydens, L., Kateman, G., Heuvel, H.M., 1993. Using genetic algorithms for an artificial neural network model inversion. Chemometrics Intell. Labor. Syst. 20 (1), 45–55.

Douglas, G., Burns, W., Bence, A., Page, S., Boehm, P., 2004. Optimizing detection limits for the analysis of petroleum hydrocarbons in complex environmental samples. Environ. Sci. Technol. 38 (14), 3958–3964.

DPR, Department of Petroleum Resources, 2002. Environmental Guidelines & Standardsfor the Petroleum Industry in Nigeria (EGASPIN). Lagos, Nigeria.

Eijsackers, H., Reinecke, A., Reinecke, S., Maboeta, M., 2017. Threatened southern African soils: A need for appropriate ecotoxicological risk assessment. Environ. Impact. Assess. Rev. 63, 128–135.

Ewenla, A., Anifowose, F.A., Akanbi, L.A., Oluwatope, O.A., Aderounmu, G.A, 2013. Prediction of Porosity and Permeability of Oil and Gas Reservoirs using Support Vector Machines and Artificial Neural Networks: A Comparative Study. In: Zelinka, Ivan, Oo, Zeya, Barsoum, Nader (Eds.), Proceedings of the 7th Global Conference on Power Control and Optimization. Prague (Czech Republic) and Yangon (Myanmar), 2008, pp. 60–66. Volume number27–28 August and 2-3 December.

FEPA, Federal Environmental Protection Agency, 1991. Guidelines and Standards for Environmental Pollution Control in Nigeria. Lagos, p. 238.

Forkuor, G., Hounkpatin, O.K.L., Welp, G., Thiel, M., 2017. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. PLoS. One 12 (1), e0170478.

George, J.K, Kumar, S., Hole, R.M., 2021. Geospatial modelling of soil erosion and risk assessment in Indian Himalayan region—A study of Uttarakhand state. Environ. Adv. 4 https://doi.org/10.1016/j.envadv.2021.100039.

Gillespie, G.D., McDonnell, K.P., O'Hare, G.M.P, 2021. Can machine learning classification Methods improve the prediction of leaf wetness in North-Western Europe compared to established empirical methods? Expert. Syst. Appl. 182, 115255.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur. J. Epidemiol. 31, 337–350. https://doi.org/10.1007/s10654-016-0149-3.

He, H., Fan, Y., 2021. A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction. Expert. Syst. Appl. 176, 114899.

Hengl, T., Leenaars, J.G.B., Shepherd, K.D., Walsh, M.G., Heuvelink, G.B.M., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., Wheeler, I., Kwabena, N.A., 2017. Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. Nutr. Cycl. Agroecosyst. 109, 77–102.

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

Htike, K.K., 2016. Efficient determination of the number of weak learners in AdaBoost. J. Exper. Theor. Artif. Intell. 29 (5), 1–16. https://doi.org/10.1080/0952813X.2016.1266038.

Huang, C., Jia, X., Zhang, Z., 2018. A Modified Back propagation artificial neural network model based on genetic algorithm to predict the flow behavior of 5754 aluminum alloy. Materials. (Basel) 11 (855), 1–15.

Jones, P.R., 2019. A note on detecting statistical outliers in psychophysical data. Atten. Percept. Psychophys. 81 (5), 1189–1196. https://doi.org/10.3758/s13414-019-01726-3.
Kazemi, S.M., Hosseini, S.M., 2011. Comparison of spatial interpolation methods for estimating heavy metals in sediments of Caspian Sea. Expert. Syst. Appl. 38, 1632–1649.
Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., Lausen, B., 2019. Ensemble of optimal trees, random forest and random projection ensemble classification. Adv. Data Anal. Classif. 1–20. https://doi.org/10.1007/s11634-019-00364-9.
Kim, K., Park, J., 2009. A Survey of applications of artificial intelligence algorithms in eco-environmental modelling. Environ. Eng. Res. 14 (2), 102–110.
Kim, S.W., Moon, J., Jeong, SW., An, Y-J., 2018. Development of a nematode offspring-counting assay for rapid and simple soil toxicity assessment. Environ. Pollut. 236, 91–99.
Kocsis, L., György, A., Andrea, N., 2013. Bán, BoostingTree: parallel selection of weak learners in boosting, with application to ranking. Mach. Learn. 93 (2–3), 293–320.
Lai, C.S., Zhong, C., Pan, K., Ng, W.W.Y., Lai, L.L., 2021. A deep learning based hybrid method for hourly solar radiation forecasting. Expert. Syst. Appl. 177, 114941.
Le Goff, L., Blot, F., Peltier, A., Laffont, L., Becerra, S., Ruiz, C.H., Abarzua, J.Q., Philippe, M., Paegelow, M., Menjot, L., Delplace, G., Schreck, E, 2022. From uncertainty to environmental impacts: reflection on the threats to water in Chacabuco Province (Chile): a combined approach in social sciences and geochemistry. Sustain. Sci. 17, 2113–2131. https://doi.org/10.1007/s11625-022-01127-w.
Li, H., Li, M., Zhao, D., Li, J., Li, S., Juhasz, A., Basta, N., Luo, Y., Ma, L., 2019. Oral Bioavailability of As, Pb, and Cd in Contaminated Soils, Dust, and Foods based on Animal Bioassays: A Review. Environ. Sci. Technol 53 (18), 10545–10559.
Li, X., Wang, H., Gu, B., 2015. Data Sparseness in Linear SVM. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), pp. 3268–3634.
Liberda, E.N., Zuk, A.M., Di, D.S., Moriarity, R.J., Martin, I.D., Tsuji, L.J.S., 2021. Complex environmental contaminant mixtures and their associations with thyroid hormones using supervised and unsupervised ML techniques. Environ. Adv. 4 https://doi.org/10.1016/j.envadv.2021.100054.
Lin, J.M., Billa, L., 2021. Spatial prediction of flood-prone areas using geographically weighted regression. Environ. Adv. 6, 100118.
Liu, W., Zhang, X., Wen, Y., Anastasio, M.A., Irudayaraj, J., 2023. A machine learning approach to elucidating PFOS-induced alterations of repressive epigenetic marks in kidney cancer cells with single-cell imaging. Environ. Adv. 11 https://doi.org/10.1016/j.envadv.2023.100344.
Matsui, T., Suzuki, K., Ando, K., Kitai, Y., Haga, C., Masuhara, N., Kawakubo, S., 2022. A natural language processing model for supporting sustainable development goals: translating semantics, visualizing nexus, and connecting stakeholders. Sustain. Sci. 17, 969–985. https://doi.org/10.1007/s11625-022-01093-3.
Mojid, M.A., Hossain, A.B.M.Z., Ashraf, M.A., 2019. Artificial neural network model to predict transport parameters of reactive solutes from basic soil properties. Environ. Pollut. 255, 113355.
Nag, U., Srivastava, S.C., Pandey, V., Kumar, S., 2005. Evolving artificial neural network with the use of hybrid strength pareto evolutionary algorithm and back propagation algorithm. In: Proceedings of the International Symposium on Intelligence Based Materials and Manufacturing. Mesra, Ranchi, India. at Birla Institute of Technnology.
NCP/BPE, 2008. Environmental Audit of Pipelines and Products Marketing Company Limited (PPMC) commissioned by the National Council on Privatization (NCP)/ Bureau of Public Enterprise (BPE). Federal Government of Nigeria: Abuja.
Nosova, A.O., Uspenskaya, M.V., 2023. Ecotoxicological effects and detection features of polyvinyl chloride microplastics in soils: A review. Environ. Adv. 13. doi.org/10.1016/j.envadv.2023.100437.
Ogunba, O.A., 2004. EIA systems in Nigeria: evolution, current practice and shortcomings. Environ. Impact. Assess. Rev. 24 (6), 643–660.
Ozigis, M.S., Kaduk, J.D., Jarvis, C.H., 2019. Mapping terrestrial oil spill impact using machine learning random forest and Landsat 8 OLI imagery: a case site within the Niger delta region of Nigeria. Environ. Sci. Pollut. Res. 26, 3621–3635.
Padarian, J., Minasny, B., McBratney, A.B., 2020. Machine learning and soil sciences: a review aided by machine learning tools. SOIL 6, 35–52.
Pandey, M., Sharma, V., 2013. A Decision tree algorithm pertaining to the student performance analysis and prediction. Int. J. Comput. Appl. 61 (13), 1–5. https://doi.org/10.5120/9985-4822.
Papageorgiou, E.I., Markinos, A., Gemptos, T., 2009. Application of fuzzy cognitive maps for cotton yield management in precision farming. Expert. Syst. Appl. 36, 12399–12413.
Pascoe, G., Riley, M., Floyd, T., Gould, C., 1998. Use of a risk-based hydrogeologic model to set remedial goals for PCBs, PAHs, and TPH in soils during redevelopment of an industrial site. Environ. Sci. Technol. 32 (6), 813–820.
Pennington, D., Ebert-Uphoff, I., Freed, N., Martin, J., Pierce, S.A., 2020. Bridging sustainability science, earth science, and data science through interdisciplinary education. Sustain. Sci. 15, 647–661. https://doi.org/10.1007/s11625-019-00735-3.
Perboli, G., Arabnezhad, E., 2021. A Machine Learning-based DSS for mid and long-term company crisis prediction. Expert. Syst. Appl. 174, 114758.
Pinedo, J., Ibáñez, R., Lijzen, JPA, Irabien, A, 2013. Assessment of soil pollution based on total petroleum hydrocarbons and individual oil substances. J. Environ. Manage 130 (30), 72–79.
Priddy, K.L., Keller, P.E., 2005. Artificial Neural Networks: An Introduction. SPIE Press, p. 180.
Pritchard, O., Hallett, S.H., Farewell, T.S., 2013. Soil corrosivity in the UK – impacts on critical infrastructure. Infrastructure Transitions Research Consortium, Working paper series. National Soil Resources Institute, Cranfield University.
Rossiter, D.G., 2018. Past, present & future of information technology in pedometrics. Geoderma 324, 131–137.
Saad, E.W., Wunsch, D.C., 2007. Neural network explanation using inversion. Neural Networks 20 (1), 78–93.
Sarma, K.K., 2018. Matlab: Demystified Basic Concepts and Applications. Vikas Publishing House, p. 200.
Segal, M.R., 2004. Machine Learning Benchmarks and Random Forest Regression, 18. Kluwer Academic Publishers. Printed in the Netherlands, pp. 1–14 peer-reviewed. https://escholarship.org/content/qt35x3v9t4/qt35x3v9t4.pdf?t=krnmaw.
Seifi, M.R., Alimardani, R., Sharifi, A., 2010. How can soil electrical conductivity measurements control soil pollution? Res. J. Environ. Earth Sci. 2 (4), 235–238.
Sezer, A., 2011. Prediction of shear development in clean sands by use of particle shape information and artificial neural networks. Expert. Syst. Appl. 38, 5603–5613.
Shapire, R., Freund, Y., Bartlett, P., Lee, W., 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. Ann. Stat. 26 (5), 1651–1686.
Singh, M., Sharma, S., Kaur, A., 2013. Performance analysis of decision trees. Int. J. Comput. Appl. 71 (19), 10–14.
Six, J., Feller, C., Denef, K., Ogle, S., de Moraes Sa, J.C., Albrecht, A., 2002. Soil organic matter, biota and aggregation in temperate and tropical soils - Effects of no-tillage. Agronomie EDP Sci. 22 (7-8), 755–775.
Song, H., Qin, A.K., Salim, F.D., 2020. Evolutionary model construction for electricity consumption prediction. Neural Comput. Appl. 32, 12155–12172.
Stenchly, K., Dao, J., Lompo, D.J.P., Buerkert, A., 2017. Effects of waste water irrigation on soil properties and soil fauna of spinach fields in a West African urban vegetable production system. Environ. Pollut. 222, 58–63.
Tamal, M., Alshammari, M., Alabdullah, M., Hourani, R., Alola, H.A, Hegazi, T.M., 2021. An integrated framework with machine learning and radiomics for accurate and rapid early diagnosis of COVID-19 from Chest X-ray. Expert. Syst. Appl. 180, 115152.
Teng, Y., Wu, J., Lu, S., Wang, Y., Jiao, X., Song, L., 2014. Soil and soil environmental quality monitoring in China: A review. Environ. Int. 69, 177–199.
Ugochukwu, C., Ochonogor, A., Jidere, C., Agu, C., Nkoloagu, F., Ewoh, J., Okwu-Delunzu, V., 2018. Exposure risks to polycyclic aromatic hydrocarbons by humans and livestock (cattle) due to hydrocarbon spill from petroleum products in Niger-delta wetland. Environ. Int. 115, 38–47.
Varjani, S., Upasani, V.N., 2019. Influence of abiotic factors, natural attenuation, bioaugmentation and nutrient supplementation on bioremediation of petroleum crude contaminated agricultural soil. J. Environ. Manage 245, 358–366.
Wang, L., Michoel, T., 2019. Accurate wisdom of the crowd from unsupervised dimension reduction. R. Soc. open sci. https://doi.org/10.1098/rsos.181806.
Wang, X., Yuan, W., Lin, C., Zhang, L., Zhang, H., Feng, X., 2019. Climate and vegetation as primary drivers for global mercury storage in surface soil. Environ. Sci. Technol. 53 (18), 10665–10675.
Wang, Y., Du, Y., Wang, J., Li, T., 2019. Calibration of a low-cost PM2.5 monitor using a random forest model. Environ. Int. 133 (Part A), 105161 https://doi.org/10.1016/j.envint.2019.105161.
Wu, Q., Wang, H., Yan, X., Liu, X., 2019. MapReduce-based adaptive random forest algorithm for multi-label classification. Neural Comput. Appl. 31, 8239–8252.
Wu, Y., 2019. Research on feature point extraction and matching machine learning method based on light field imaging. Neural Comput. Appl. 31, 8157–8169.
Xu, Q., Deng, K., Jiang, C., Sun, F., Huang, X., 2017. Composite quantile regression neural network with applications. Expert. Syst. Appl. 76, 129–139.
Yang, J., Wang, J., Li, A., Li, G., Zhang, F., 2020. Disturbance, carbon physicochemical structure, and soil microenvironment codetermine soil organic carbon stability in oilfields. Environ. Int. 135, 105390 https://doi.org/10.1016/j.envint.2019.105390.
Yang, L., Liu, S., Tsoka, S., Papageorgiou, L.G., 2017. A regression tree approach using mathematical programming. Expert. Syst. Appl. 78, 347–357.
Yin, C., 2021. International law regulation of offshore oil and gas exploitation. Environ. Impact. Assess. Rev. 88, 106551 https://doi.org/10.1016/j.eiar.2021.106551.
Zeng, Q., Liu, Y., Zhao, H., Sun, M., Li, X., 2017. Comparison of models for predicting the changes in phytoplankton community composition in the receiving water system of an inter-basin water transfer project. Environ. Pollut. 223, 676–684.
Zhang, D., Yan, D., Cheng, H., Fang, W., Huang, B., Wang, X., Wang, X., Yan, Y., Ouyang, C., Li, Y., Wang, Q., Cao, A., 2020. Effects of multi-year biofumigation on soil bacterial and fungal communities and strawberry yield. Environ. Pollut. 256, 113415.
Zhu, Y., Tan, Y., Hua, Y., Wang, M., Zhang, G., Zhang, J., 2010. Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. J. Digit. Imaging 23 (1), 51–65. https://doi.org/10.1007/s10278-009-9185-9.