© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1720339

Artificial Intelligence-Based Prediction of Bioremediation of Crude Oil-Contaminated Soil

GAIUS ILIYA[1*], ABDULSALAM SURAJUDEEN[2], KABIRU IBRAHIM MUSA[3]
[1, 2, 3]*Center for Embedded Artificial Intelligence and Smart Energy Systems, Faculty of Engineering and Engineering Technology, Abubakar Tafawa Balewa University, Gubi Campus, Bauchi, Nigeria.*
**Corresponding Author*

Abstract- Crude oil contamination of soil remains a critical environmental challenge, particularly across petroleum-producing regions of the developing world. Although biological remediation strategies have attracted considerable research interest over the past decade, the empirical monitoring of remediation processes is constrained by the labor-intensive nature of laboratory measurements and the absence of reliable predictive frameworks. This work introduces a data-driven, machine-learning framework for forecasting bioremediation efficiency (BE%) in crude oil-contaminated soil, using a dataset of 14 monitored field samples measured fortnightly across an eight-week remediation cycle. To overcome the data scarcity challenge, a kinetic spline augmentation protocol expanded the original 56-observation dataset to 2,114 temporally dense records. Six supervised regression algorithms, namely Linear, Ridge, and Lasso Regression, Support Vector Regression, Gradient Boosting Regression, and a Multilayer Perceptron Artificial Neural Network (MLP-ANN), were trained and assessed using a five-fold cross-validation scheme. Among the six architectures tested, the MLP-ANN delivered the strongest predictive accuracy, returning a coefficient of determination (R^2) of 0.9803, an adjusted R^2 of 0.9802, a root-mean-square error (RMSE) of 2.9613%, a mean absolute error (MAE) of 2.0970%, and a mean absolute percentage error (MAPE) of 4.29%, surpassing every other model evaluated. When deployed as a binary classifier at a regulatory compliance threshold of BE% of at least 70%, the ANN achieved a classification accuracy of 96.45%. SHapley Additive exPlanations (SHAP) analysis identified oil and grease concentration as the dominant predictor of remediation outcome, contributing a mean absolute SHAP value of 16.89, followed by moisture content (1.33), microbial count (1.11), and elapsed treatment time (0.39). Residual diagnostic analyses confirmed approximate normality and homoscedasticity of model errors. These findings establish the ANN as a robust, interpretable, and practically deployable tool for real-time monitoring and decision support in soil bioremediation programmes.

Index Terms- bioremediation efficiency, artificial neural network, machine learning, crude oil contamination, SHAP interpretability, soil remediation prediction.

I. INTRODUCTION

Petroleum hydrocarbon contamination of terrestrial soils constitutes one of the foremost threats to ecosystem stability, food security, and public health in oil-producing nations. The Niger Delta region of Nigeria, among other petroleum-rich regions globally, has borne disproportionate ecological damage from upstream spills, pipeline vandalism, and inadequate remediation practice [1], [2]. Crude oil seeps into the soil, disrupts its natural structure, surpresses microbial diversity, reduces soil fertility, and introduces polycyclic aromatic hydrocarbons (PAHs) and other hazardous constituents that persist for years without deliberate intervention [3], [4].

Biological remediation, which harnesses the metabolic activity of indigenous and augmented microorganisms to mineralize hydrocarbon contaminants, now stands as one of the most environmentally sustainable and economically practical solutions for site restoration [7], [8], [9]. Strategies including biostimulation, bioaugmentation, phytoremediation, and vermiremediation have been evaluated extensively across laboratory and field scales [10], [11], [12]. Organic amendments such as cow dung, poultry manure, soybean hull, and composted materials have been shown to supply the nitrogen and phosphorus necessary to sustain high microbial activity and accelerate total petroleum hydrocarbon (TPH) reduction [13], [14], [15].

Despite the breadth of experimental bioremediation literature, a critical methodological gap persists: the

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1720339

ability to predict bioremediation efficiency dynamically from readily measurable soil and environmental parameters without requiring repeated costly laboratory analyses. Traditional empirical monitoring is temporally sparse, dependent on specialist equipment, and incapable of real-time decision support. Kinetic models such as the Monod and Haldane equations have provided valuable understanding of the underlying processes [16], but their predictive utility is constrained by strong parametric assumptions and limited applicability across diverse site conditions.

Recent progress in machine learning offers a methodologically robust alternative. Artificial Neural Networks (ANN) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) have been successfully applied to model TPH degradation [17], [18], [19], predict phytotoxicity from soil chemistry [20], and optimize bioremediation parameters in slurry bioreactor systems [21]. More broadly, artificial intelligence tools are being integrated into soil remediation research across systematic review, biopile design, and field-scale decision support frameworks [3], [22], [23].

This work adds to the expanding body of machine-learning applications within environmental remediation in three distinct ways. First, it applies a kinetic spline augmentation protocol to generate a temporally dense training dataset from compact field-monitored bioremediation data, a methodological approach that receives limited attention in the current literature. Second, it provides a systematic six-model comparative evaluation under five-fold cross-validation, yielding reliable performance estimates across multiple complementary metrics. Third, it integrates SHAP interpretability analysis to identify the relative contribution of each predictor variable to bioremediation efficiency, addressing the practical communication gap between predictive models and environmental practitioners. The outcomes are intended to support evidence-based, data-driven management of contaminated site remediation in both research and applied contexts.

II. RELATED WORK

2.1. Crude Oil Contamination and Bioremediation Strategies

Crude oil is a heterogeneous composition of aliphatic hydrocarbons, cycloalkanes, aromatic compounds, and polar fractions including resins and asphaltenes. When introduced into the soil environment, it disrupts soil structure, impairs aeration, and creates hydrophobic barriers that reduce water infiltration and suppress seed germination [5], [25]. The extent of contamination damage is dose-dependent, with high oil concentrations producing near-complete inhibition of plant growth and microbial activity in early exposure stages [6].

Bioremediation exploits the capacity of microorganisms to catabolize hydrocarbon compounds, converting them into carbon dioxide, water, and biomass under aerobic conditions [1], [4]. Indigenous hydrocarbon-utilising bacteria (HUB) and fungi (HUF), particularly strains of Pseudomonas, Bacillus, Enterobacter, Rhodococcus, and Aspergillus, constitute the primary degradation agents in most contaminated terrestrial systems [26], [27], [28]. The performance of these organisms is governed by substrate bioavailability, nutrient balance, moisture content, temperature, and pH [29], [30].

Biostimulation with organic amendments introduces limiting nutrients and creates favourable physiochemical conditions for degrading microorganisms. Studies using cow dung [8], [13], [14], poultry manure [11], pigeon droppings [15], composted agricultural wastes [31], and soybean hull [32], [33] have reported total hydrocarbon content reductions ranging from 33% to over 99% across experimental periods of two to twelve weeks. Bioaugmentation with carefully selected microbial consortia further enhances degradation where indigenous populations are poorly adapted to the contaminant profile [9], [34].

Phytoremediation and vermiremediation represent additional biological strategies with utility in field applications. Water hyacinth, water lettuce, Costus afer, and biochar-assisted plant growth promotion have been explored with promising results [24], [35], [36], [37]. Earthworms, particularly Lumbricus

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1720339

terrestris, have been found to enhance soil aeration, nutrient cycling, and hydrocarbon bioavailability through vermiremediation [12], [19]. Reviews of integrated approaches show that combining two or more biological strategies consistently outperforms single-agent interventions [2], [10].

2.2. Machine Learning in Bioremediation Prediction
The application of machine learning to bioremediation prediction addresses a fundamental limitation of purely empirical approaches: the inability to interpolate or forecast degradation trajectories from limited monitoring data. [17] reported that an ANFIS model achieved $R^2 = 0.9349$ in predicting TPH degradation from composting experiments, outperforming a conventional ANN model ($R^2 = 0.8862$). [20] applied both ANN and Support Vector Regression to predict phytotoxicity of petroleum-contaminated soils, with SVR achieving a coefficient of determination of 0.80 using soil physicochemical parameters as inputs.

[18] deployed MATLAB-based ANN modelling to predict residual hydrocarbon content in chicken droppings-amended soils, identifying the Scaled Conjugate Gradient algorithm as optimal and achieving $R^2 = 0.998$. [21] applied ANN to model n-hexadecane removal in slurry bioreactors, finding that bacterial consortia improved model predictive correlation compared to pure culture inputs. [19] paired response surface methodology with both ANN and ANFIS to forecast crude-oil removal achieved via vermicompost treatment, recording removal efficiencies ranging from 29% to 98.9%, with the ANFIS model attaining a correlation coefficient of 0.9978 in the column-remediation trial.

Recent literature shows AI techniques and bioremediation research beginning to intersect more closely. [3] used AI-powered literature mapping tools to systematically review 59 studies on PAH and heavy metal remediation. [22] reviewed a decade of biopile-based remediation advances, including the integration of AI optimisation tools. [23] specifically examined AI applications in oil and gas site decommissioning and remediation. Despite these advances, there remains a notable gap in the literature combining kinetic spline augmentation, comprehensive multi-model benchmarking, and SHAP interpretability within a unified framework applied to a field-monitored bioremediation dataset, which the present work seeks to address.

III. MATERIALS AND METHODS

3.1 Experimental Dataset
The raw dataset was obtained from a monitored bioremediation experiment conducted across 14 soil sample sites (S1 to S14) [40], with measurements recorded at fortnightly intervals at weeks 2, 4, 6, and 8 of the treatment period. Each record consisted of six variables: a sample identifier, treatment week as the temporal predictor, oil-and-grease (O&G) concentration measured in parts per million (ppm), microbial population density expressed in colony-forming units per gram of soil (cfu/g), percentage moisture content, and bioremediation efficiency (BE%), which served as the target output. The raw dataset contained 56 observations (14 samples across 4 time points). Across the raw dataset, BE% ranged from a minimum of -16.35% in one anomalous sample at week 8 (Sample S13, indicative of re-contamination or measurement variability) to a maximum of 90.28%, with a mean of approximately 59.96% across all 56 observations. Oil and grease concentrations ranged from 9,333 ppm to 119,333 ppm, microbial counts from approximately 24.9 million to 3.71 billion cfu/g, and moisture content from 1.67% to 25.39%.

3.2 Kinetic Spline Data Augmentation
Given the small size of the raw experimental dataset (n = 56), direct application of high-capacity machine learning models would yield unreliable cross-validated estimates due to insufficient sample diversity. To address this, a kinetic spline augmentation protocol was designed to generate a temporally dense representation of bioremediation pathways while preserving the empirical relationships observed in the raw data.

For each of the 14 sample groups, the four measured time points were used as anchor nodes for piecewise linear interpolation (numpy.interp) across a dense grid of treatment weeks covering 2.0 to 8.0 at intervals of 0.04 weeks, producing 151 interpolated time steps per sample. Zero-mean Gaussian noise, scaled to 1.5% of each feature's own standard deviation, was injected into every interpolated value. This introduced a

(1)

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1720339

controlled degree of randomness that avoided exact duplication of records while keeping the dataset's original distributional shape intact. The augmentation procedure is summarised as:

$$\hat{X}(w) = \text{interp}(w; w_0, x_0) + \varepsilon, \quad \varepsilon \sim N(0, 0.015^2\sigma^2_{\text{feature}})$$

In Eq. (1), $\hat{x}(w)$ denotes the interpolated/augmented feature value at treatment week w, w_0 the vector of measured anchor time points, x_0 the corresponding measured feature values at those anchors, *interp (.)* the piecewise linear interpolation function, ε the added Gaussian noise term, and $\sigma_{feature}$ the standard deviation of the respective feature computed from the raw dataset, with noise scaled to 1.5% of that standard deviation. Following interpolation, boundary clipping was applied to ensure physical realism: BE% values were clipped to the interval [−20%, 100%], while O&G concentration, microbial count, and moisture content were constrained to non-negative values with a lower bound of 0.1. This procedure expanded the dataset from 56 to 2,114 observations (14 samples × 151 time steps), representing a 37.75-fold increase in data volume. The augmented dataset preserved the mean bioremediation efficiency of 63.89% (standard deviation 21.09%), mean O&G concentration of 27,925 ppm, mean microbial count of 5.45×10^8 cfu/g, and mean moisture content of 9.93%.

3.3 Study Process Flowchart

With the raw field measurements now expanded into a temporally dense and physically realistic dataset, it is useful to step back and view the study as a single coherent pipeline before the remaining methodological steps are detailed individually. Fig. 1 traces this pipeline from the raw dataset and kinetic spline augmentation described above, through feature engineering, model training and cross-validation, to the interpretability and diagnostic checks presented later in the paper.

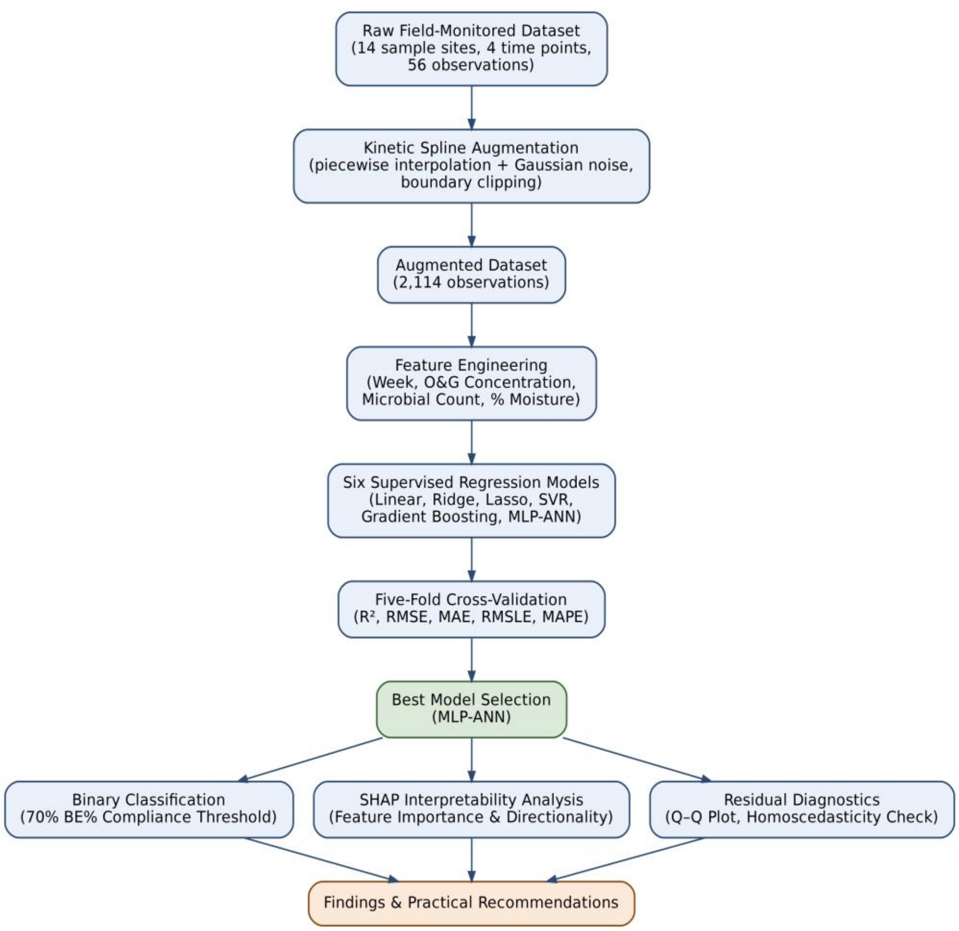

Fig 1. Overall Process Flowchart of the Study

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1720339

3.4 Feature Engineering and Model Inputs

Four input features were used in all models: treatment Week, oil and grease concentration (O&G Concentration, ppm), microbial count (Microbial Count, cfu/g), and percentage moisture content (% Moisture Content). The target variable was bioremediation efficiency (BE%). The Pearson correlation matrix computed on the augmented dataset revealed that O&G concentration exhibited the strongest negative linear association with BE% ($r = -0.872$), pointing to its dominant role as a predictor. Treatment week showed a moderate positive correlation ($r = 0.561$), while microbial count and moisture content exhibited weaker associations ($r = 0.413$ and $r = -0.136$, respectively). No evidence of harmful multicollinearity was detected between the predictor variables.

3.5 Machine Learning Models

Six supervised regression architectures were evaluated. Linear Regression established a parametric baseline under ordinary least squares assumptions. Ridge Regression (alpha = 1.0) introduced L2 regularization to penalize coefficient magnitude, reducing variance without feature elimination. Lasso Regression (alpha = 0.1) applied L1 regularization to encourage sparsity in the coefficient vector. The Support Vector Regressor (SVR) employed a radial-basis-function kernel, with inputs standardized beforehand through a preprocessing pipeline. The Gradient Boosting Regressor (GBR) employed an ensemble of 100 sequential decision trees with a fixed random seed. All models involving regularization or kernel methods were embedded within scikit-learn pipeline objects that applied StandardScaler transformation as a preprocessing step.

The best-performing model, the Multilayer Perceptron Artificial Neural Network (MLP-ANN), comprised three fully connected hidden layers containing 128, 64, and 32 neurons in turn, each using the Rectified Linear Unit (ReLU) activation function. The Adam optimizer was used with an adaptive learning rate schedule, L2 regularization parameter alpha = 0.001, and a maximum iteration budget of 1,500 epochs. The network architecture is described by the forward propagation equations:

$$H^{(1)} = \mathrm{ReLU}(W^{(1)}x + b^{(1)}) \quad (2)$$

$$h^{(2)} = \mathrm{ReLU}(W^{(2)}h^{(1)} + b^{(2)}) \quad (3)$$

$$h^{(3)} = \mathrm{ReLU}(W^{(3)}h^{(2)} + b^{(3)}) \quad (4)$$

$$\hat{y} = W^{(4)}h^{(3)} + b^{(4)} \quad (5)$$

$$\mathrm{ReLU}(z) = \max(0, z) \quad (6)$$

in which W^l and b^l represent, respectively, the weight matrix and bias vector belonging to layer l, with the ReLU function defined as max(0, z).

3.6 Model Evaluation Protocol

Every model underwent five-fold cross-validation (configured with KFold, shuffle = True, random_state = 42), and the cross_val_predict routine was used to obtain out-of-fold predictions covering the full augmented dataset. This approach provides unbiased performance estimates without information leakage between training and evaluation subsets. Six quantitative metrics were computed for each model:

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i|$$

$$RMSLE = \sqrt{\frac{1}{n}\sum(\log(1 + y_i) - \log((1 + \hat{y}_i))^2} \quad (7)$$

$$MAPE = \frac{100}{n}\sum\frac{|y_i - \hat{y}_i|}{y_i} \quad (8)$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, v\ where\ SS_{res} = \sum(y_i - \hat{y}_i)^2, \quad (9)$$

$$SS_{tot} = \sum(y_i - \bar{y})^2 \quad (10)$$

$$R^2{}_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-p-1} \quad (11)$$

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (12)$$

Here, n denotes the total number of observations, p the count of predictor variables, y_i the measured values, $\hat{y}_i$ the corresponding predicted values, and $\bar{y}$ the average of the observed values.

3.7 Binary Classification and SHAP Analysis

To assess operational utility in a regulatory context, the continuous ANN predictions and the observed BE% values were discretized at a threshold of 70%, where values at or above this threshold were classified as High Compliance (bioremediation target achieved) and values below were classified as Low Yield. A confusion matrix was computed, and classification accuracy was reported. This threshold aligns with the efficiency levels frequently cited in experimental literature as indicative of successful remediation under organic amendment management practices [8], [14].

Shapley Additive exPlanations (SHAP) values were generated using the TreeExplainer method, applied to a Gradient Boosting model trained on the complete augmented dataset. Rooted in cooperative game theory, SHAP breaks each individual prediction down into additive contributions from each feature, yielding both an overall ranking of feature importance (mean |SHAP|) and feature-level directional patterns visualized in the beeswarm plot. This interpretability layer ensures that model outputs can be communicated to practitioners and policymakers in a physically meaningful form.

3.8 Residual Diagnostics

The quality of ANN predictions was assessed through two standard regression diagnostic procedures. A normal quantile–quantile (Q–Q) plot was generated by comparing the empirical quantiles of the out-of-fold ANN residuals against the quantiles expected under a theoretical Gaussian distribution. Substantial alignment along the reference line indicates approximate normality of the error distribution. An error dispersion plot was constructed by plotting residuals against fitted values to assess homoscedasticity, that is, the constancy of residual variance across the range of predictions. Systematic patterns in this plot would indicate heteroscedastic errors requiring corrective transformation.

IV. RESULTS AND DISCUSSION

4.1 Target Variable Distribution and Predictor Profiles

The distribution of bioremediation efficiency across the 2,114 augmented observations is presented in Fig. 2. The density profile displays a left-skewed pattern with the majority of observations concentrated in the 70% to 90% BE% range, consistent with the progressive nature of microbial hydrocarbon degradation over the eight-week monitoring period. The ideal Gaussian reference line indicates that the target variable departs from strict normality, a characteristic common in bounded environmental efficiency metrics. The observed skewness is consistent with bioremediation dynamics in which early weeks record low efficiency that rises steeply as microbial populations establish themselves, as documented in some studies [13], [38].

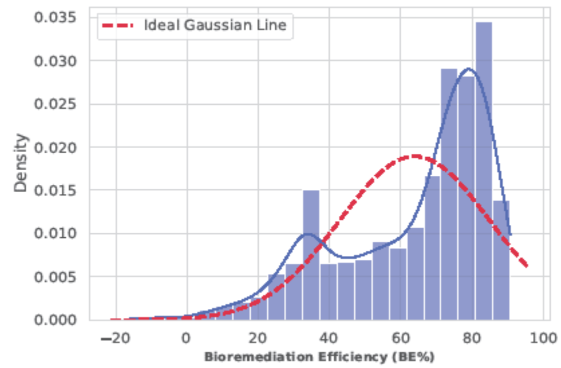

Fig 2. Target Variable Density and Normality Profile Analysis of Bioremediation Efficiency (BE%).

The predictor distribution profiles illustrated in Fig. 3 reveal that O&G concentration and microbial count exhibit substantially wider ranges and several extreme observations relative to treatment week and moisture content. Microbial count in particular, varies over three orders of magnitude across the augmented dataset (24.9×10^6 to 3.71×10^9 cfu/g), a pattern consistent with the accelerated growth of hydrocarbon-utilising bacterial consortia following biostimulant addition documented across multiple studies [15], [39]. The box plot shows that all four predictors contain detectable but not disqualifying outliers, supporting the suitability of the augmented dataset for machine learning analysis.

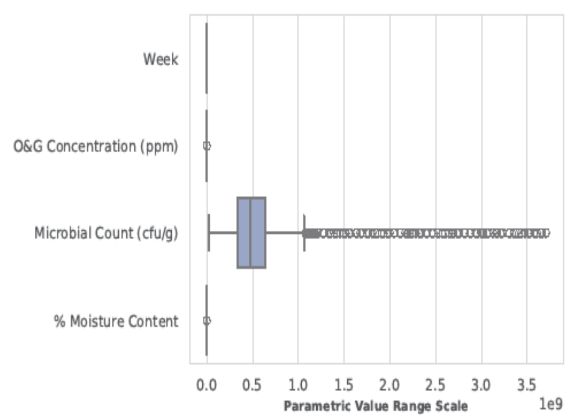

Fig 3. Horizontal Box-and-Whisker Predictor Distribution and Outlier Screening Profile.

4.2 Correlation Structure and Feature Relationships

The Pearson correlation topography heatmap in Fig. 4 quantifies the bivariate linear relationships among all five variables in the augmented dataset. The dominant correlation in the matrix is the strongly negative association between O&G concentration and BE% (r = −0.872), consistent with the intuitive, physically expected relationship that declining hydrocarbon concentration corresponds to advancing bioremediation efficiency. Treatment week shows a moderate positive correlation with BE% (r = 0.561), in line with efficiency improving over time as microbial communities adapt and grow. Microbial count is positively associated with BE% (r = 0.413), consistent with the well-documented relationship between population density of hydrocarbon-utilising bacteria and degradation rate [26], [27].

Moisture content exhibits a weak negative correlation with BE% (r = −0.136), which may be explained by the higher initial moisture levels observed in less contaminated samples, rather than any suppressive effect of moisture on degradation. The moderate negative correlation between treatment week and O&G concentration (r = −0.406) is consistent with hydrocarbon levels declining over the course of the experiment, as expected. No predictor pair exceeds a correlation of 0.44 in absolute terms, indicating an absence of severe multicollinearity. These correlation findings align with SHAP results discussed in Section 4.6.

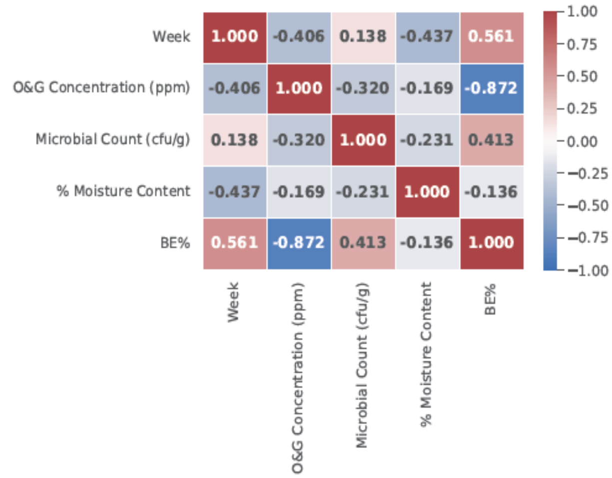

Fig 4. Multi-Variable Pearson Correlation Topography Heatmap.

4.3 Comparative Model Performance

Table 1 presents the full six-metric evaluation matrix for all six machine learning architectures, ranked by R^2 in descending order. The results reveal a clear performance hierarchy across the model families, with the MLP-ANN achieving the highest performance on all six metrics and the three linear regression variants occupying the lower performance band.

Table 1. Cross-Validated Multi-Metric Performance Comparison of Six Machine Learning Models

Model Architecture	R^2	Adj. R^2	RMSE	MAE	RMSLE	MAPE (%)
Artificial Neural Network (MLP)	0.9803	0.9802	2.9613	2.0970	0.0732	4.2856
Gradient Boosting Regressor	0.9480	0.9479	4.8066	3.6472	0.1411	8.7860
Support Vector Regressor	0.8807	0.8805	7.2840	4.9000	0.3088	18.1842
Ridge Regression	0.8535	0.8532	8.0721	6.5091	0.3636	21.5585
Linear Regression	0.8535	0.8532	8.0721	6.5090	0.3636	21.5587
Lasso Regression	0.8534	0.8531	8.0737	6.5102	0.3624	21.6106

The ANN achieved an R^2 of 0.9803, indicating that the model accounts for 98.03% of variance in bioremediation efficiency across the augmented dataset. Its RMSE of 2.9613% and MAE of 2.0970% are operationally negligible for a target variable covering a range of approximately 110 percentage points. The Gradient Boosting Regressor ranked second with R^2 = 0.9480 and RMSE = 4.8066%, representing a 38% increase in prediction error relative to the ANN. The SVR achieved R^2 = 0.8807 with an RMSE of 7.2840%, while all three linear models clustered around R^2 = 0.853 and RMSE of approximately 8.07%, exposing the limitations of linear assumptions in capturing non-linear biodegradation kinetics.

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1720339

The MAPE of 4.29% for the ANN compares favourably with the 8.79% of the gradient boosting model and the 21.56% typical of linear regressors. The RMSLE of 0.0732 for the ANN points to low relative error on a log scale, important given the orders-of-magnitude variation in microbial count. These results are consistent with prior findings in bioremediation modelling literature: [17] reported that their best-performing ANFIS achieved RMSE values well below those of standard ANN architectures, and [21] found that consortium-based ANN models outperformed pure-culture approaches. The superiority of the MLP-ANN over gradient boosting in this study may stem from the model's capacity to learn smooth interpolation of kinetic degradation curves that gradient boosting, being a piecewise-constant ensemble method, approximates less precisely.

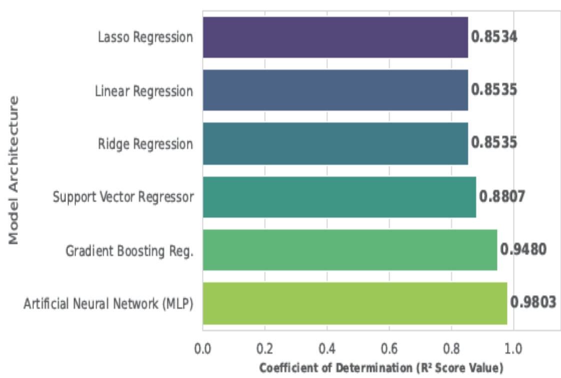

Fig 5. Cross-Validated Model Evaluation R^2 Performance Comparison Bar Chart.

4.4 ANN Training Convergence and Prediction Accuracy

Figure 6 illustrates the training loss trajectory of the MLP-ANN over successive optimization epochs when trained on an 80/20 hold-out split with early stopping enabled. The loss curve descends steeply in the initial epochs, driven by rapid gradient-based adjustment of the network weights, before transitioning to a slow asymptotic decline. Convergence occurred comfortably within the 600-epoch training allowance, a result of the combined benefit of the Adam optimizer, an adaptive learning-rate schedule, and L2 weight decay applied for regularization. The smooth, monotonically decreasing curve without notable oscillations indicates stable optimization without overfitting to the training partition.

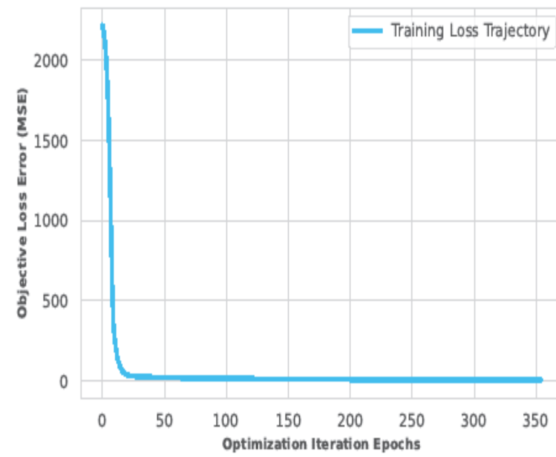

Fig 6. ANN Training Loss Convergence Curve (Objective Loss Error per Epoch)

The scatter plot in Figure 7 depicts ANN out-of-fold predictions against observed BE% values across all 2,114 augmented observations. Points are tightly clustered around the crimson identity line (representing perfect prediction) with minimal systematic deviation across the full range of the target variable. A marginal widening of prediction variance is observable in the mid-range of BE% values (20% to 50%), corresponding to the transition phase of bioremediation where microbial population dynamics are most volatile. This residual uncertainty is physically expected and does not represent a systematic model failure.

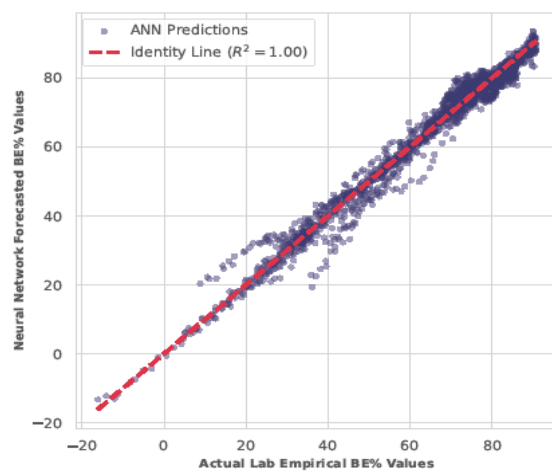

Fig 7. ANN Predictions versus Actual BE% Values Scatter Plot

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1720339

4.5 Binary Classification Performance

When the ANN was applied as a binary classifier at the 70% BE% compliance threshold, the resulting confusion matrix (Fig. 8) shows that the model correctly identified 922 of 972 Low Yield instances (94.86% recall) and 1,117 of 1,142 High Compliance instances (97.81% recall). The overall classification accuracy of 96.45% across 2,114 predictions shows that the ANN can reliably distinguish remediating soil systems that have reached an operationally satisfactory level of hydrocarbon removal from those that have not. The 75 misclassified instances (50 false positives and 25 false negatives) represent boundary cases clustered near the 70% threshold, where small prediction errors of a few percentage points determine the classification outcome.

This level of binary classification accuracy carries direct practical significance. Environmental regulators and site managers who deploy this model for remediation monitoring can expect fewer than 4 incorrect classification decisions in 100 assessments, reducing the risk of premature site closure or unnecessary extension of expensive treatment programmes. The threshold of 70% is consistent with the efficiency levels at which several experimental studies have declared bioremediation interventions successful [8], [14], [29].

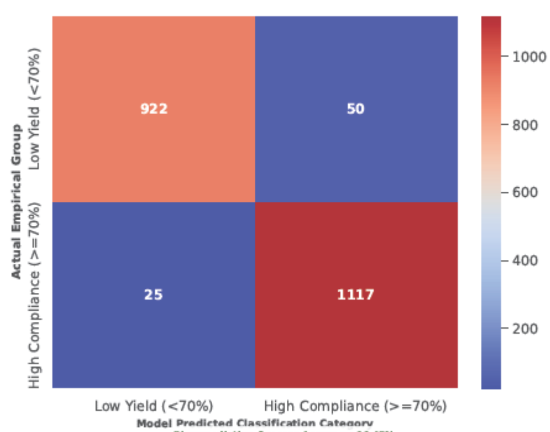

Fig. 8. Confusion Matrix for Binary Bioremediation Success Classification (Threshold: BE% >= 70%)

4.6 SHAP Interpretability Analysis

The SHAP global importance bar chart (Fig. 9) reveals that O&G concentration is overwhelmingly the dominant predictor of bioremediation efficiency, with a mean absolute SHAP value of 16.89, which is more than twelve times larger than the contribution of the second-ranked predictor, moisture content (1.33). Microbial count ranked third (1.11) and treatment week fourth (0.39).

The SHAP beeswarm plot (Fig. 10) provides additional directional information. High O&G concentration values (depicted in red) push model predictions strongly in the negative direction, showing that elevated hydrocarbon loading suppresses predicted BE%. Low O&G concentrations (blue) are associated with large positive SHAP contributions, consistent with the high BE% values observed in later weeks when hydrocarbon concentrations have declined substantially. This pattern is corroborated by the Pearson correlation result ($r = -0.872$) and by the broader experimental literature documenting the dose-dependent suppression of microbial efficiency at high petroleum concentrations [6], [16].

The modest SHAP contributions of microbial count and moisture content suggest that, while these variables carry predictive information, their direct effects are substantially mediated by the trajectory of hydrocarbon concentration over time. Treatment week's low SHAP contribution does not imply that time is unimportant; rather, this shows that the temporal signal is largely captured by the correlated decline in O&G concentration, making week partially redundant as a standalone predictor once O&G is included. These findings have a direct implication for practitioners: reliable real-time monitoring of bioremediation progress may be achievable from hydrocarbon concentration measurements alone, with microbial and moisture parameters serving as useful secondary indicators.

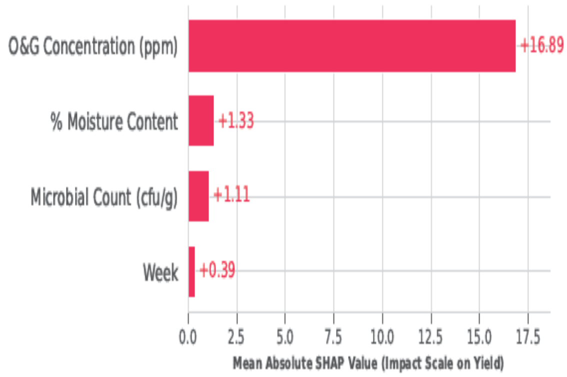

Fig 9. SHAP Global Feature Importance Ranking Bar Chart

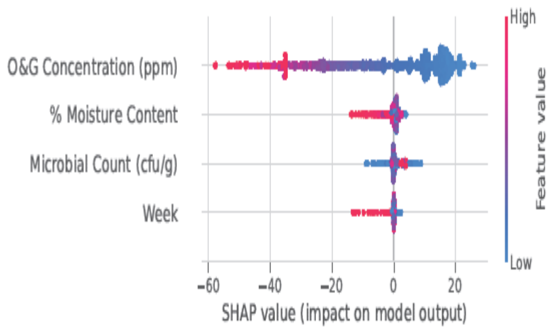

Fig. 10 SHAP Summary Beeswarm Plot Showing Directional Feature Impact Topologies

4.7 Residual Diagnostics

Figure 11 presents the Q-Q plot of ANN out-of-fold residuals. The residual points trace closely along the theoretical normal reference line across the central range of quantiles, indicating approximate Gaussian distribution of prediction errors. Moderate deviations are visible in the tails, particularly for quantiles beyond +/-2 standard deviations, suggesting the presence of a small number of heavy-tailed outlier residuals. These tail deviations correspond to the anomalous samples noted in the raw dataset, especially Sample S13, which exhibited a negative BE% value in week 8 attributable to possible re-contamination or field measurement uncertainty. The central body of the residual distribution is sufficiently normal to support the validity of the model's parametric inference.

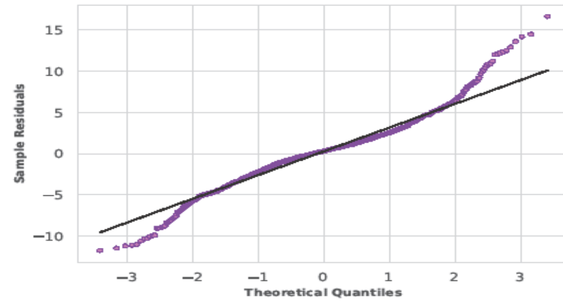

Fig 11. Regression Residual Normal Quantile-Quantile (Q-Q) Diagnostic Verification

The error dispersion plot in Figure 12 shows residuals plotted against ANN-fitted values. Residuals are distributed approximately symmetrically around the zero reference line across the full prediction range of -20% to 90% BE%, with no systematic fan-shaped expansion of variance that would indicate heteroscedasticity. A modest cluster of slightly larger positive residuals appears in the prediction range of 20% to 40%, corresponding to the volatile mid-phase of bioremediation trajectories. Overall, the dispersion pattern is broadly consistent with homoscedastic errors, supporting the reliability of cross-validated RMSE as a meaningful performance measure. These diagnostic findings are consistent with ANN residual patterns observed in analogous studies [19], [16].

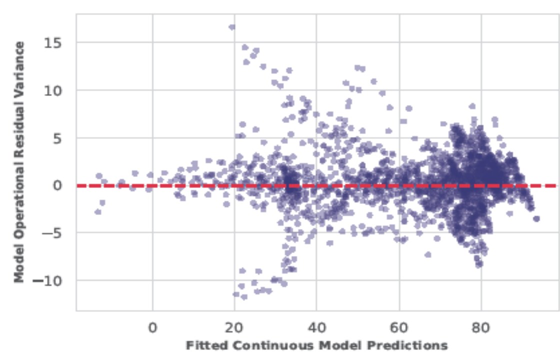

Fig 12. Error Variance Dispersion Homoscedasticity Regression Diagnostic Plot.

4.8 Comparison with Prior Literature

The MLP-ANN's R^2 of 0.9803 in this study is comparable to or exceeds the performance of machine learning models reported in analogous bioremediation prediction tasks. [18] achieved $R^2 = 0.998$ but on a much smaller dataset with potential overfitting concerns. [17] reported ANFIS $R^2 = 0.9349$ on a constrained experimental design. [20] reported SVR $R^2 = 0.80$ for phytotoxicity prediction. The current study's combination of 2,114 augmented observations, five-fold cross-validation, and multi-model benchmarking provides a methodologically robust performance estimate relative to these precedents. The 96.45% classification accuracy represents an evaluation approach that receives limited coverage in the existing bioremediation AI literature, where binary compliance classification from continuous efficiency predictions has seldom been formally assessed.

4.9 Future Research Directions

Building on the results above, this framework could be extended to include additional site-specific variables such as soil pH, nutrient concentrations (N and P), contaminant speciation, and temperature, which have been identified as important drivers of bioremediation kinetics in the experimental literature [6], [29], [30]. Application of deep recurrent architectures such as Long Short-Term Memory networks may further improve temporal prediction by explicitly capturing the sequential dynamics of microbial degradation. Validation of the current model on independently collected bioremediation datasets from diverse geographic contexts and amendment treatment protocols would further strengthen its generalizability as a practitioner-grade predictive tool.

V. CONCUSION

This study has shown that machine learning, and in particular the MLP-ANN, can predict bioremediation efficiency in crude oil-contaminated soils with high accuracy from a small set of routinely measurable environmental parameters. The kinetic spline augmentation framework successfully addressed the data scarcity challenge inherent to field-monitored bioremediation experiments, enabling reliable cross-validated model evaluation. The neural network attained an R^2 value of 0.9803 alongside an RMSE of 2.9613%, exceeding the performance of the other five models tested, and yielded a 96.45% binary classification accuracy when assessed against a 70% bioremediation-efficiency compliance threshold.

SHAP analysis established that oil and grease concentration is by far the most influential predictor of remediation outcome, followed at a distance by moisture content, microbial count, and treatment week. Residual diagnostics confirmed approximate normality and homoscedasticity of ANN prediction errors, lending statistical credibility to the performance metrics reported. These findings advance the scientific case for deploying data-driven predictive models as operational tools in contaminated site management, offering the prospect of significantly reduced monitoring costs and more timely remediation decisions.

Author Contributions

All authors contributed significantly to the conception and design of the study, and to the acquisition, analysis, and interpretation of data. All authors participated in drafting the article and revising it critically for important intellectual content, and approved the final version for publication.

Funding

This research received no dedicated funding from any public, commercial, or not-for-profit funding agency.

Data Availability Statement

The data used in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest relating to the research, authorship, or publication of this work.

REFERENCES

[1] B. A. Mekonnen, T. A. Aragaw, and M. B. Genet, "Bioremediation of petroleum hydrocarbon contaminated soil: a review on principles, degradation mechanisms, and advancements," Frontiers in Environmental Science, vol. 12, p. 1354422, 2024, doi: 10.3389/fenvs.2024.1354422.

[2] I. E. Miriogu, U. I. Uchendu, H. C. Nwanekezi, and E. M. Okoro, "Remediating crude oil polluted sites using integrated bioremediation technologies: a review," Nigerian Journal of Environmental Sciences and Technology, vol. 8, no. 2, pp. 97–116, 2024.

[3] Z. Ashkanani, R. Mohtar, S. Al-Enezi, P. K. Smith, S. Calabrese, X. Ma, and M. Abdullah, "AI-assisted systematic review on remediation of contaminated soils with PAHs and heavy metals," Journal of Hazardous Materials, vol. 468, p. 133813, 2024, doi: 10.1016/j.jhazmat.2024.133813.

[4] I. M. S. Anekwe and Y. M. Isa, "Bioremediation of crude oil-contaminated soils: a review," Petroleum and Coal, vol. 64, no. 3, pp. 665–681, 2022.

[5] O. M. Agbogidi and O. Ogbe, "Performance of cassava (Manihot esculenta Crantz) as affected by crude oil contamination of soil," Journal of Science Research and Reviews, vol. 2, no. 2, pp. 63–69, 2025.

[6] C. M. Camacho-Montealegre, E. M. Rodrigues, D. K. Morais, and M. R. Totola, "Prokaryotic community diversity during bioremediation of crude oil contaminated oilfield soil: effects of hydrocarbon concentration and salinity," Brazilian Journal of Microbiology, vol. 52, no. 2, pp. 787–800, 2021.

[7] S. Ramesh, "Utilizing bacteria for crude oil-contaminated soil bioremediation and monitoring through tomato plant growth," Nature Environment and Pollution Technology, vol. 24, no. 2, 2025.

[8] C. E. Okafor, O. P. Nwwabueze, C. P. Uzuegbu, S. C. Okeke, and R. C. Okarfor, "Bioremediation efficacy and total petroleum hydrocarbon reduction in crude oil contaminated soil using cow dung," Journal of Applied Sciences and Environmental Management, vol. 29, no. 2, pp. 555–561, 2025.

[9] S. Hosseini, R. Sharifi, and A. Habibi, "Efficient bioremediation of crude oil contaminated soil by a consortium of in-situ biosurfactant producing hydrocarbon-degraders," Scientific Reports, vol. 15, no. 1, p. 19852, 2025, doi: 10.1038/s41598-025-05035-8.

[10] A. B. Stanojevic, M. Vrvic, J. Szakova, and S. Miletic, "Evaluation of the ex-situ bioremediation of the petroleum hydrocarbons contaminated soil," Bioremediation Journal, vol. 28, no. 4, pp. 553–563, 2024, doi: 10.1080/10889868.2023.2283580.

[11] R. K. Douglas, P. P. Araka, A. Fou, and A. Hart, "Evaluation of the potential of agricultural wastes-cattle manure and poultry manure for bioremediation of crude oil-contaminated soil," Bioremediation Journal, vol. 29, no. 1, pp. 96–103, 2025.

[12] B. M. Younus, M. S. Alenazi, and M. M. Akba, "Bioremediation of petroleum-contaminated soil using earthworms: a study on hydrocarbon reduction in Iraq," Journal of Education for Pure Science, vol. 15, no. 1, 2025.

[13] O. E. Oludele, M. E. Wyse, O. K. Odeniyi, P. O. Ali, and M. Kugbogbenmowei, "Bioremediation of crude oil contaminated soil using cow dung," International Journal of Sciences, vol. 10, no. 1, pp. 37–45, 2021.

[14] J. C. Ogbu, G. Etuk-Udo, O. C. Prince, and D. Luka, "Bioremediation of crude oil contaminated agricultural soil using cow dung," Journal of Infectious Diseases and Treatment, vol. 3, no. 2, pp. 1–4, 2025.

[15] J. O. Williams, R. R. Nrior, R. Renner, and E. J. Nkpornwi, "Synergistic bioremediation of crude oil contaminated soil: role of bioreactor and pigeon droppings," Journal of Advances in Microbiology Research, vol. 6, no. 1, pp. 23–33, 2025.

[16] G. Menon and A. Mane, "Analysis of diesel spill bioremediation by bacterial isolates in batch culture using Monod and Haldane kinetic model," Journal of Advanced Scientific Research, vol. 12, no. 03 Suppl 2, pp. 270–275, 2021.

[17] K. A. Ani and C. M. Agu, "Predictive comparison and assessment of ANFIS and ANN, as efficient tools in modeling degradation of total petroleum hydrocarbon (TPH)," Cleaner Waste Systems, vol. 3, p. 100052, 2022, doi: 10.1016/j.clwas.2022.100052.

[18] S. E. Uwadiae and C. D. Obasi, "Modelling of bioremediation of oil-contaminated soil using chicken droppings as biostimulant," Journal of Applied Sciences and Environmental Management, vol. 25, no. 11, pp. 1895–1898, 2021.

[19] P. O. Ehiomogue, I. I. Ahuchaogu, and U. I. Udoumoh, "Remediation of crude oil-

contaminated soil using vermicompost in the Niger Delta area of Nigeria," Arid Zone Journal of Engineering, Technology and Environment, vol. 19, no. 4, pp. 719–732, 2023.

[20] D. Shadrin, M. Pukalchik, E. Kovaleva, and M. Fedorov, "Artificial intelligence models to predict acute phytotoxicity in petroleum contaminated soils," Ecotoxicology and Environmental Safety, vol. 194, p. 110410, 2020.

[21] R. Morovati, F. Abbasi, M. R. Samaei, H. Mehrazmay, and A. R. Lari, "Modelling of n-hexadecane bioremediation from soil by slurry bioreactors using artificial neural network method," Scientific Reports, vol. 12, no. 1, p. 19662, 2022, doi: 10.1038/s41598-022-21996-6.

[22] C. Di Marcantonio, L. Passatore, M. Zacchini et al., "Advancements in biopile-based sustainable soil remediation: a decade of improvements, integrating bioremediation technologies and AI-based innovative tools," Environmental Science and Pollution Research, 2025, doi: 10.1007/s11356-025-37002-1.

[23] I. K. Egbuna, H. Agboro, O. O. Nwachukwu, F. E. George, J. B. Asere, and S. A. Ogunkanmi, "Artificial intelligence for predictive analysis, efficiency improvement and reduction in carbon footprint during decommissioning and site remediation in oil and gas fields," World Journal of Advanced Research and Reviews, vol. 26, no. 2, pp. 3394–3405, 2025.

[24] M. H. Ali, M. T. Sattar, M. I. Khan, M. Naveed, M. Rafique, S. Alamri, and M. H. Siddiqui, "Enhanced growth of mungbean and remediation of petroleum hydrocarbons by Enterobacter sp. MN17 and biochar addition in diesel contaminated soil," Applied Sciences, vol. 10, no. 23, p. 8548, 2020.

[25] W. A. Ali, W. A. Farid, and A. N. K. Al-Salman, "Bioremediation of agricultural soil contaminated by a crude oil spill," Applied Ecology and Environmental Research, vol. 18, no. 1, 2020.

[26] R. Babalola, V. E. Efeovbokhan, Y. O. Atiku, U. E. Usoro, M. A. Ibeh, E. E. Alagbe, and O. G. Abatan, "Slurry-phase bioremediation of Ogoni land crude oil contaminated soil," IOP Conference Series: Materials Science and Engineering, vol. 1107, no. 1, p. 012167, 2021.

[27] P. Behdarvandan, R. Jalilzadeh Yengejeh, S. Sabzalipour, L. Roomiani, and K. Payandeh, "Bioremediation of crude oil by indigenous species isolated from oil sludge contaminated soil: a case study of Karun Gas Oil Production Company, Iran," Journal of Advances in Environmental Health Research, vol. 8, no. 4, pp. 234–241, 2020.

[28] M. Ejaz, B. Zhao, X. Wang, S. Bashir, F. U. Haider, Z. Aslam, M. I. Khan, M. Shabaan, M. Naveed, and A. Mustafa, "Isolation and characterization of oil-degrading Enterobacter sp. from naturally hydrocarbon-contaminated soils and their potential use against the bioremediation of crude oil," Applied Sciences, vol. 11, no. 8, p. 3504, 2021.

[29] W. H. Tsai, T. W. Chen, Y. H. Liu, S. M. Shen, C. S. Chen, and C. J. Tien, "A field-scale assessment of the efficacy of bioremediation of petroleum hydrocarbon-contaminated soils using two biostimulants," International Biodeterioration and Biodegradation, vol. 196, p. 105942, 2025, doi: 10.1016/j.ibiod.2024.105942.

[30] E. Atai, R. B. Jumbo, T. Cowley, I. Azuazu, F. Coulon, and M. Pawlett, "Efficacy of bioamendments in reducing the influence of salinity on the bioremediation of oil-contaminated soil," Science of the Total Environment, vol. 892, p. 164720, 2023.

[31] P. E. Edekor and S. E. Uwadiae, "Bioremediation of crude oil-contaminated soil using compost as bio-stimulant," Journal of Applied Sciences and Environmental Management, vol. 25, no. 11, pp. 1855–1858, 2021.

[32] O. K. Jimoh-Hamza and A. T. Ajao, "Bioremediation of crude oil-polluted soil with a consortium of Pseudomonas aeruginosa and soybean hull," UNIOSUN Journal of Engineering and Environmental Sciences, vol. 6, no. 2, pp. 1–10, 2024.

[33] A. T. Ajao and O. K. Jimoh-Hamza, "Exploring the effectiveness of Pseudomonas aeruginosa isolates for bioremediation of crude oil-contaminated soils using soybean hull as a

biostimulant: a focus on ETPH and PAHs," Fudma Journal of Sciences, vol. 8, no. 4, pp. 296–302, 2024.

[34] T. V. Funtikova, L. I. Akhmetov, I. F. Puntus, P. A. Mikhailov, N. O. Appazov, R. A. Narmanova, A. E. Filonov, and I. P. Solyanikova, "Bioremediation of oil-contaminated soil of the Republic of Kazakhstan using a new biopreparation," Microorganisms, vol. 11, no. 2, p. 522, 2023.

[35] G. E. Agu, "Bioremediation of a crude oil contaminated soil using water hyacinth (Eichhornia crassipes)," World Journal of Advanced Research and Reviews, vol. 18, no. 3, pp. 880–888, 2023.

[36] G. E. Agu and A. I. Hart, "Bioremediation of a crude oil contaminated soil using water lettuce (Pistia stratiotes)," Scientia Africana, vol. 22, no. 2, pp. 167–176, 2023.

[37] C. Emeka, R. N. Okparanma, and O. W. Achinike, "Bioremediation of total polycyclic aromatic hydrocarbon in crude oil-contaminated soil using Costus afer plant," Research Journal of Ecology and Environmental Sciences, vol. 3, no. 1, pp. 26–36, 2023.

[38] F. G. Evans, U. H. Nkalo, D. Amachree, and M. O. Raimi, "From killer to solution: evaluating bioremediation strategies on microbial diversity in crude oil-contaminated soil over three to six months in Port Harcourt, Nigeria," Advances in Environmental and Engineering Research, vol. 5, no. 4, pp. 1–26, 2024.

[39] M. C. Maduwuba, "Bacteriological and physicochemical analysis of a crude oil-polluted soil undergoing laboratory-scale bioremediation," Animal Research International, vol. 21, no. 2, pp. 5443–5452, 2024.

[40] Atiku, Y. M., "Biostimulation of soil contaminated with spent motor oil using cow dung and poultry litter in land farming microcosm," *IOSR Journal of Environmental Science, Toxicology and Food Technology*, vol. 12, no. 3, pp. 1–10, 2018.