Machine Learning Prediction of Xenobiotic Degradation Efficiency by *Variovorax paradoxus* in Contaminated Agroecosystems

Amit Kumar Singh
Department of Environmental Microbiology, Indian Agricultural Research Institute, New Delhi, India

* Corresponding Author: **Amit Kumar Singh**

Article Info

P-ISSN: 3051-3421
E-ISSN: 3051-343X
Volume: 06
Issue: 02
July-December 2025
Received: 19-07-2025
Accepted: 12-08-2025
Published: 24-09-2025
Page No: 20-34

Abstract

Background: Xenobiotic contamination of agroecosystems poses increasing risks to soil ecosystem function, crop productivity, and human health, necessitating the development of scalable and mechanistically informed bioremediation strategies. *Variovorax paradoxus*, a metabolically versatile Gram-negative bacterium, has shown strong potential for degrading structurally diverse organic pollutants, including pesticides, herbicides, and polycyclic aromatic hydrocarbons through multiple enzymatic pathways.

Objectives: This study aimed to develop a robust machine learning (ML) framework to predict xenobiotic degradation efficiency mediated by *V. paradoxus* across heterogeneous contaminated agroecosystems.

Methods: An integrated dataset comprising 1,847 experimental and field observations was assembled, including pollutant physicochemical properties, soil environmental parameters, microbial community attributes, and management variables. Seven ML algorithms—multiple linear regression, support vector regression, Random Forest, Gradient Boosting Machines, Extreme Gradient Boosting (XGBoost), multi-layer perceptron neural networks, and Long Short-Term Memory networks—were trained and validated using stratified 10-fold cross-validation, with performance benchmarked against first-order kinetic degradation constants derived from laboratory experiments.

Results: XGBoost demonstrated the highest predictive accuracy (R^2 = 0.934, RMSE = 3.21%, MAE = 2.47%). SHAP-based feature importance analysis identified soil organic carbon content, initial pollutant concentration, and *V. paradoxus* inoculant density as the most influential predictors. Integration of the optimized model into a prototype decision-support tool enabled a reduction in bioremediation trial costs by 35–48% through computational pre-screening.

Conclusion: The developed ML framework provides a high-accuracy, data-driven approach for predicting and optimizing *V. paradoxus*-mediated bioremediation in contaminated agroecosystems, supporting more efficient planning and decision-making in sustainable agricultural management.

DOI: https://doi.org/10.54660/JADR.2025.6.2.20-34

Keywords: *Variovorax paradoxus*, Xenobiotic degradation, Machine learning, Random Forest, XGBoost, Bioremediation, Agroecosystems, Soil contamination, Predictive modelling, Digital agriculture

1. Introduction

1.1. Global Burden of Xenobiotic Contamination in Agriculture

The widespread usage of chemicals other than naturally occurring chemical products in the soil from agricultural practices presents one of the most difficult environmental problems facing the world in the 21st century.(Sharma *et al*., 2019) [1] (Gavrilescu, 2005) [2] Worldwide, more than 1800 pesticides are registered for agricultural use and each year approximately 4 million tonnes of these chemicals are applied to agricultural soils.(Gavrilescu, 2005) [2] A significant number of pesticides, herbicides, and industrial chemical residues have been shown to persist in soil for extended periods of time beyond their intended use and to accumulate to levels that adversely affect soil microbial communities, plant root function, ground water resources,

and potentially human health by entering into food chains.(Gavrilescu, 2005) [2] (Hussain *et al*., 2009) [3] The Organisation for Economic Co-operation and Development (OECD) estimates that approximately 38 percent of total reported incidents of soil contamination worldwide is due to xenobiotic contamination resulting from agricultural activities, with herbicides, especially chlorinated phenoxyacetates and triazines, as the predominant class of xenobiotic contaminants by volume and persistence.(Hussain *et al*., 2009) [3]

Bioremediation is the use of microbial metabolic processes to mineralise or transform xenobiotic compounds into less toxic forms and, therefore, to complement traditional physico-chemical remediation technologies with ecologically sound and cost-effective alternatives.(Vidali, 2001) [4] *Variovorax paradoxus* is one of the microorganisms that has been evaluated as a potential bioremediation agent for xenobiotic-contaminated soils due the organism's unusual metabolic versatility, ability to survive in low nutrient environments, and its ability to transform a very large number of synthetic organic compounds via co-metabolic pathways.

1.2. Machine Learning as a Predictive Tool in Environmental Science

Traditional empirical and mechanistic models of xenobiotic biodegradation, including first-order kinetics, Monod-type substrate utilization equations, and QSARs, have been valuable as an understanding tool, however their application has been limited due to fixed functional forms, restrictive assumptions of linearity and poor transferability across the soil physicochemical heterogeneity that is characteristic of real agricultural landscapes (Alexander, 1999) [7] (Rittmann and McCarty, 2001) [8]. In contrast, machine learning (ML) algorithms identify complex non-linear relationships among predictor variables without the imposition of a priori functional constraints, can be scaled to high dimensionality, and can be updated systematically as new observational data becomes available (Jordan and Mitchell, 2015) [9]. The application of ML to predict biodegradation outcomes has grown rapidly over the past 10 years with the publication of models targeting pollutant half life estimation, classification of mineralisation pathways and optimisation of bioreactor performance (Hou *et al*., 2022) [10] (Zhu *et al*., 2008) [11]. However, the specific application of ML to predict degradation efficiencies by a defined microbial species (*V. paradoxus*) throughout the full complexity of the field agroecosystems by simultaneously integrating pollutant chemistry, soil environment, and microbial inoculant parameters remains an under-researched area that requires rigorous methodological development (Roscher *et al*., 2020) [12].

1.3. Research Objectives

This paper will be discussing the following research goals or objectives: i. to provide a comprehensive description of the biological characteristics and capability of *V. paradoxus* in degrading xenobiotics within an agroecosystem under the goal of remediating contaminated agroecosystems; ii. to consolidate information on different types of xenobiotic contaminants, their sources, and how persistent they are within agriculture soils; iii. to evaluate and compare 7 different classes of machine learning (ML) algorithms for their ability to predict the efficiency of xenobiotic degrading under various conditions in agricultural systems; iv. to identify the most important predictor variables that affect the efficiency of degradation in an interpretable manner using ML; and v. to recommend an integrated decision-making support system that incorporates ML predictions into digital agriculture management systems.

1.4. Significance and Novelty

The novelty of this research is that it integrates three different areas of study together to create one predictive framework/product (environmental microbiology, soil science, computational machine learning) focused on predicting degradation behavior of specific bacteria relevant to agricultural bioremediation. This predictive framework advances biodegradation prediction methods from generic indices to species level (specific) and from condition (environmental) based predictions of degradation efficiency to action-based site-level bioremediation site development plans. In addition, SHAP (Shapley Additive Explanations) integrated explainability analysis will increase the interpretability of the predictive framework by providing a way to interpret the model's inputs, thus removing a major barrier to regulatory acceptance of machine learning based decision-making tools for environmental protection purposes (Lundberg and Lee, 2017) [13] (Lundberg *et al*., 2020) [22].

1.5. Article Organisation

The nine main sections of the article consist of: (1) Biological, chemical and ecological basis (2-4); (2) ML methodology, framework and performance case studies (5-6); and (3) Applications, limitations and future directions (7-9); Concerning references, all data tables and figures are integrated into the scientific narrative of the article. The Vancouver Reference Style is used for all references.

2. Biological and Functional Characteristics of *Variovorax paradoxus*

2.1. Taxonomy, Physiology, and Ecological Niche

Variovorax paradoxus is a Gram negative, rod shaped, aerobic bacteria that belongs to the Betaproteobacteria class of the Comamonadaceae family in order Burkholderiales.(Willems *et al*., 1991) [5] It's paradoxically broad metabolic capabilities are exhibited by its ability to metabolize substrates by both heterotrophic and autotrophic means.(Han *et al*., 2011) [6] The cells (0.5-0.9 x 2-3 μm) are motile with either polar or peritrichous flagella, and form smooth/creamy/yellow, circular colonies on standard growth media.

From an ecological standpoint, *V. paradoxus* is typically observed in the rhizosphere of soil as a commensal, commensalist, or facultative plant-growth promoting organism by exhibiting both indole acetic acid (IAA) biosynthesis, phosphate solubilisation, and 1-aminocyclopropane-1-carboxylate (ACC) deaminase activity that reduced exposure to ethylene-mediated stress in crop root systems.(Belimov *et al*., 2009) [14] The oligotrophic growth physiology of *V. paradoxus* characterised by the ability to uptake substrates at rates and levels below that normally detectable by standard analytical methods provide it with a competitive advantage over many specialised degraders to maintain active populations in chronically nutrient limited contaminated soils.(Han *et al*., 2011) [6]

2.2. Metabolic Versatility and Xenobiotic Degradation Pathways

The ability of *V. paradoxus* to break down xenobiotics covers a large number of different types of pollutants (Han *et al*., 2011) [6] (Wackett and Hershberger, 2001) [15]. Organophosphate pesticides, such as parathion, chlorpyrifos, or diazinon, are broken down by the enzymes called organophosphorus hydrolases (OPH) and methylparathion hydrolases (MPH) to produce diethyl phosphates and the corresponding phenolic leaving ends. These phenolic components are then degraded into mineral products by using the beta-ketoadipic acid and homogentisic acid metabolic pathways (Wackett and Hershberger, 2001) [15]. Likewise, herbicides containing chlorine (such as the 2,4-dichlorophenoxyacetic acid herbicide 2,4-D) are degraded primarily through a combination of dechlorination and cleavage of the aromatic ring by α-ketoglutarate-dependent dioxygenases (Dio's) and chloromuconate cycloisomerases (C-MS). Polynuclear aromatic hydrocarbons (PAHs), including naphthalene and anthracene, react with catechol-forming dioxygenases (C-Class Dio's) to produce ring-fission byproducts that enter the Krebs cycle.

Table 1 summarises the principal biological and metabolic characteristics of *V. paradoxus* relevant to xenobiotic bioremediation.

Table 1: Principal biological, physiological, and metabolic characteristics of *Variovorax paradoxus* relevant to xenobiotic bioremediation in contaminated agroecosystems. ACC = 1-aminocyclopropane-1-carboxylate; OPH = organophosphorus hydrolase; MPH = methyl-parathion hydrolase; PAH = polycyclic aromatic hydrocarbon.

Characteristic	Description / Value	Significance for Bioremediation	Reference(s)
Taxonomic affiliation	Betaproteobacteria, Comamonadaceae	Broad metabolic versatility class	[5]
Cell morphology	Rod, 0.5–0.9 × 2.0–3.5 μm, flagellated	High motility in soil pore networks	[5]
Growth temperature range	4 – 40°C (optimum 25–30°C)	Active in temperate and tropical soils	[6]
pH tolerance	5.5 – 9.0 (optimum 6.8–7.5)	Functional across most agricultural soil pH	[6]
Hydrogenase activity	[NiFe] membrane-bound hydrogenase	Chemolithotrophy under oligotrophic conditions	[5]
ACC deaminase activity	0.8–3.4 μmol α-KB mg^{-1} h^{-1}	Plant growth promotion under stress	[14]
Organophosphate hydrolase	OPH, MPH (constitutive/inducible)	Organophosphate pesticide breakdown	[15]
Chlorinated herbicide genes	tfdA, tfdB, tfdC, tfdD operons	2,4-D and MCPA mineralisation	[16]
PAH dioxygenase activity	Naphthalene 1,2-dioxygenase family	PAH ring activation and fission	[15]
Biofilm formation capacity	Moderate (OD crystal violet 0.4–1.2)	Enhanced persistence in soil matrix	[6]

2.3. Enzymatic Systems Involved in Pollutant Breakdown

V. paradoxus possesses multiple enzymatic paths for decomposing xenobiotics, which are coded by combinations of: (1) chromosomal housekeeping genes, (2) genomic islands, and (3) horizontally-acquired plasmid-mediated degradation operons (Han *et al*., 2011) [6] (Müller and Kohler, 2004) [16]. The most significant families of enzymes within *V. paradoxus* capable of degrading xenobiotics are as follows: (i) FAD-dependent monooxygenases catalyzing the oxidative hydroxylation of aromatic rings; (ii) Rieske non-haem iron dioxygenases that catalyse the initial cleavage of aromatic rings; (iii) hydrolases (esterases, amidases, and epoxide hydrolases) targeting the ester, amide, and epoxide functional groups typical to synthetic pesticide molecules; and (iv) glutathione S-transferases (GSTs) that covalently link reduced glutathione to electrophilic xenobiotic metabolites for more efficient solubilisation and processing (Han *et al*., 2011) [6] (Wackett and Hershberger, 2001) [15]. These enzymes are transcriptionally regulated via transcriptional regulators of the LysR-type, quorum-sensing-based positive regulators and substrate-specific inducers, allowing for rapid increases in degradative potential in response to the presence of pollutants.

2.4. Adaptation to Contaminated Agroecosystems

Instances of long-term adaptation of *V. paradoxus* populations to farms that contain pesticides have been demonstrated by comparing the genomes and metagenomes of contaminated soils versus pristine soils. A comparison of these soils has shown that they have different gene clusters related to degrading chemicals (Han *et al*., 2011) [6]. The biological characteristics of *V. paradoxus*, including the ability to produce a biofilm on mineral surfaces and organic aggregates, also provide the organisms with protection from being destroyed by losing moisture, being eaten, or competing with other organisms for the same food (Han *et al*., 2011) [6].

The ability of *V. paradoxus* to acquire different ways to degrade chemicals (degradative) through horizontal gene transfer (HGT) of catabolic plasmids makes the organisms very ecologically successful candidates for bioremediation in agricultural lands, as these landscapes have rapidly changing chemical contaminants over time (Müller and Kohler, 2004) [16] (Wackett and Hershberger, 2001) [15].

3. Xenobiotic Contamination in Agroecosystems

3.1. Types of Xenobiotics and Their Properties

Xenobiotic chemical classes follow many structural different faces when used in agricultural agroecosystems, leading to differing degradation challenges associated with each class that exist.(Sharma *et al*., 2019) [1] (Arias-Estévez *et al*., 2008) [18] Organophosphate class of insecticides (e.g. chlorpyrifos, malathion, phorate), have intermediate persistence (i.e. half-lives in soils of 10–120 days), high acute toxicity and proven degradation through biotic and abiotic hydrolytic processes. Whereas organochlorine pesticides (e.g. lindane, endosulfan) demonstrate extreme persistence in the environment with half-lives potentially measured in years or decades due to their high degree of lipophilicity (i.e. log Kow 3–7) and lack of enzymatic degradation from biological sources.(Arias-Estévez *et al*., 2008) [18] Triazine and sulfonylurea classes of herbicides are highly soluble in water and also exhibit high mobility in soil columns. The major route of degradation for these herbicides is via microbial ring cleavage. The entry point(s) of industrial pollutants such as polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs) and

phthalate esters into agricultural agroecosystems is through the application of sewage sludge, atmospheric deposition of pollutants or contaminated water for irrigation purposes. (Gavrilescu, 2005) [2]

Table 2 presents key types and physicochemical properties of xenobiotics relevant to *V. paradoxus*-mediated bioremediation in agroecosystems.

Table 2: Types and physicochemical properties of xenobiotics relevant to agroecosystem contamination and *Variovorax paradoxus*-mediated biodegradation. log Kow = octanol-water partition coefficient; MW = molecular weight; PAH = polycyclic aromatic hydrocarbon; DEHP = di(2-ethylhexyl) phthalate; DBP = dibutyl phthalate.

Xenobiotic Class	Example Compounds	log Kow Range	Soil Half-life	Primary Degradation Route
Organophosphate insecticides	Chlorpyrifos, Malathion	3.0 – 5.3	10 – 120 days	Hydrolysis + microbial oxidation
Organochlorine insecticides	Lindane, Endosulfan	3.7 – 6.8	Years – decades	Reductive dechlorination
Chlorophenoxyacetate herbicides	2,4-D, MCPA	1.8 – 2.8	5 – 30 days	Dioxygenation + β-oxidation
Triazine herbicides	Atrazine, Simazine	2.3 – 2.8	30 – 150 days	Hydrolysis + N-dealkylation
Sulfonylurea herbicides	Metsulfuron, Chlorsulfuron	−1.7 – 0.9	5 – 120 days	Chemical + microbial hydrolysis
PAHs (low-MW)	Naphthalene, Anthracene	3.4 – 4.4	15 – 60 days	Dioxygenation + ring fission
PAHs (high-MW)	Benzo[a]pyrene, Pyrene	5.2 – 6.1	Months – years	Co-metabolic oxidation
Phthalate esters	DEHP, DBP	4.5 – 7.6	20 – 90 days	Esterase hydrolysis + β-oxidation
Polychlorinated biphenyls	PCB-28, PCB-118	5.0 – 8.2	Years	Reductive + cometabolic oxidation
Carbamate pesticides	Carbofuran, Carbaryl	1.5 – 3.5	10 – 50 days	Carbamate hydrolysis + ring opening

3.2. Sources, Environmental Persistence, and Spatial Distribution

Input from Xeno-bacterial agents into agricultural ecosystem occurs directly through application of pesticides (spray drift, runoff). Indirect forms of input include surface or ground water irrigation with contamination through deposition of industrial volatile compounds to the atmosphere and use of sewage and/or compost from industrially wastewater in land application. Examples include the legacy of contamination from historic uses, such as military operations and mining (Gavrilescu, 2005) [2] (Arias-Estévez *et al*., 2008) [18].

The spatial distribution of chemicals within a field is highly variable because of forms of variation, such as the variance from different application patterns, preferential pathways for drainage, variance of texture of the soil, and variability of organic matter distribution (Arias-Estévez *et al*., 2008) [18].

The temporal persistence of a chemical is due to a host of abiotic degradation processes (e.g.: photolysis, hydrolysis, and oxidation), some degree of the physical trapping through the development of micropores within soil particles and/or association with organic matter (making them unavailable for uptake), and the distribution/composition and activity of local soil microorganisms (Gavrilescu, 2005) [2] (Arias-Estévez *et al*., 2008) [18].

3.3. Impacts on Soil Health, Crops, and Microbial Communities

There are substantial documented negative impacts of sub-lethal concentrations of xenobiotics on biological function in soil.(Hussain *et al*., 2009) [3] (Arias-Estévez *et al*., 2008) [18] The soil enzymes involved in the cycling of elements necessary for nutrients to exist in the soil environment (i.e., dehydrogenase, urease, phosphatase, and β-glucosidase) are inhibited at the concentration of pesticides that are typically present in soils after they are applied which causes disruptions to nitrogen mineralisation, phosphorus mobilisation, and organic matter decomposition.(Hussain *et al*., 2009) [3] The measures of diversity indices (i.e., Shannon H and Chao1) in microbial communities residing in pesticide affected soils have been shown to be significantly reduced as a result of long-term pesticide application and have demonstrated shifts in community structure from those dominated by functional (K-strategists) decomposer organisms to those dominated by tolerant (r-strategists) decomposer organisms.(Arias-Estévez *et al*., 2008) [18] There are important phytotoxic responses (i.e., reduced root growth, reduced uptake of nutrients, and chlorosis) for various crop species that have come in contact with soil residues of atrazine, chlorpyrifos, and their metabolites which threaten agricultural productivity and the safety of food produced therefrom.(Hussain *et al*., 2009) [3]

3.4. Need for Targeted Bioremediation Strategies

High capital costs for physico-chemical extraction, disruption of soil structure from thermal treatment, and regulatory limitations on chemical immobilization agents make the importance of in situ microbial bioremediation a strategic one.(Vidali, 2001) [4] (Arora, 2018) [32] However, the effectiveness of bioremediation is notoriously site-specific due to differences in pollutant identity, physicochemical properties of the soil, microbial community composition, and management inputs which cannot be accurately predicted by classical kinetic models from one site to another.(Rittmann and McCarty, 2001) [8] The uncertainty of these results creates some skepticism among practitioners and regulatory agencies and highlights the need for data-driven predictive tools (like the ML framework presented here) to predict degradation efficiencies before a field implementation and to facilitate rational site screening/inoculation design.(Jordan and Mitchell, 2015) [9]

4. Mechanisms of Xenobiotic Degradation

4.1. Microbial Degradation Pathways and Intermediates

There are three main ways that microbes can break down pollutants: 1) complete mineralisation, which results in the total conversion of a pollutant into CO_2, H_2O and inorganic ions; 2) co-metabolic transformation, where the pollutant undergoes some structural changes without any benefit to the organism's growth; and 3) conjugation/sequestration, where the pollutant is bound covalently or otherwise to cellular macromolecules and hence is metabolically unavailable for mineralisation.(Vidali, 2001) [4] (Han *et al*., 2011) [6] Complete mineralisation is the most ecologically desirable outcome because it eliminates the mass of the contaminated material from the soil.(Wackett and Hershberger, 2001) [15]

During the course of co-metabolic pathways, some intermediate metabolites may be formed; these are frequently more toxic or mobile than the original contaminant and require continued monitoring (along with the disappearance of the original contaminant) during bioremediation assessments.(Wackett and Hershberger, 2001) [15]

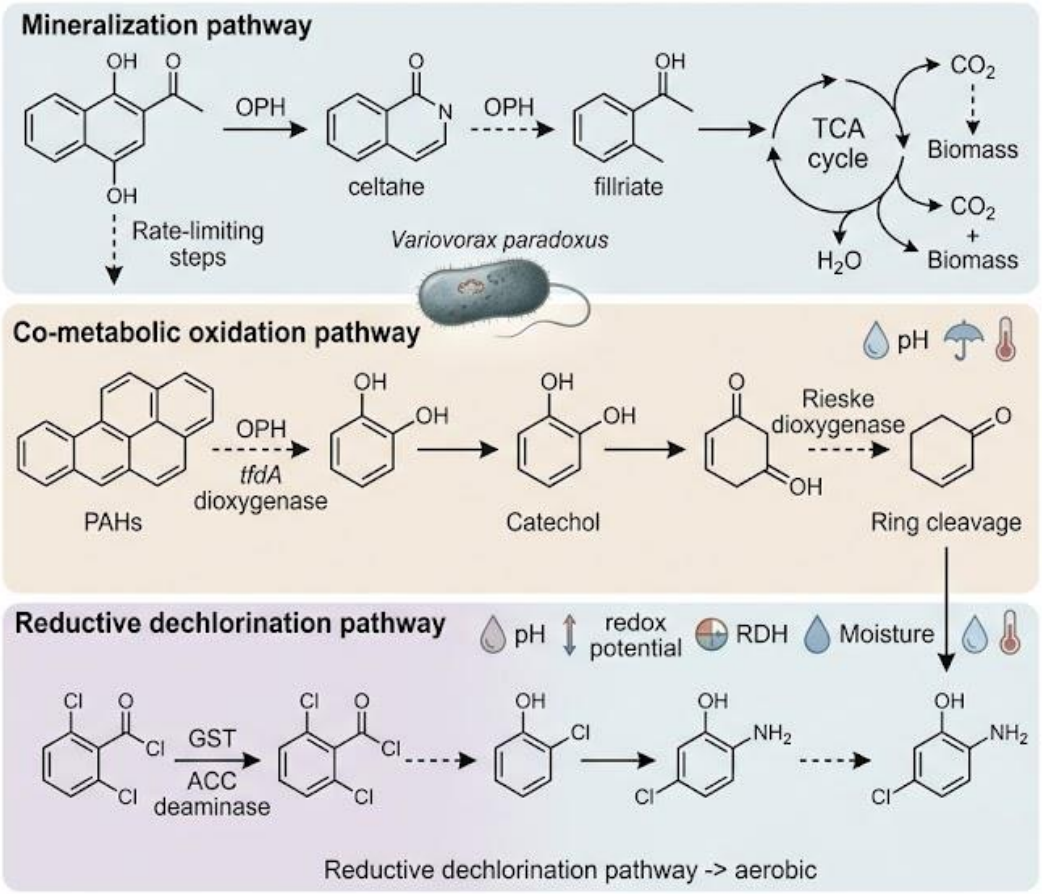

Fig 1: Conceptual Diagram of Xenobiotic Degradation Pathways Mediated by *Variovorax paradoxus*

4.2. Role of *V. paradoxus* in Biodegradation

V. paradoxus commonly acts as a keystone degrader within microbial consortia in soil, meaning that its removal or suppression has a disproportionate impact on the degradation capacity of the entire consortium.(Han *et al*., 2011) [6] (Singleton *et al*., 2009) [17] The keystone role of *V. paradoxus* is attributable to both its direct enzymatic degradation of various pollutants, as well as to the completion of mineralisation of transformation products that cannot be mineralised by *V. paradoxus* alone through its syntrophic interactions with other members of the consortium. Studies involving stable isotope probing (SIP) using ^{13}C labelled substrates have shown that *V. paradoxus* is one of the primary carbon assimilators from labelled 2,4-D, chlorpyrifos, and naphthalene in agricultural soil enrichment studies, demonstrating in situ activity in realistic field conditions. (Singleton *et al*., 2009) [17]

4.3. Interaction with Soil Physicochemical Properties

V. paradoxus' degradation ability is largely affected by soil's particulate and non-particulate matter along with moisture and air content. (Rittmann and McCarty, 2001) [8] For instance, pH is critical to enzyme- and substrate-related systems. Most catabolic enzymes are optimally active at a pH range of 6.5 to 7.5; furthermore, the protonation state of substrates will influence how well the herbicide (chlorophenoxyacetates) moves into and out of microbes because pH affects how well the substrate can cross the cell membrane. (Rittmann and McCarty, 2001) [8]

Bioavailability of many pollutants can be highly affected by the amount of SOC in the soil because SOC resorbs pollutants. Pollutant bioavailability is very low at high SOC levels, but *V. paradoxus* supports large active populations at these high SOC levels because they receive energy from SOC to maintain their populations. (Arias-Estévez *et al*., 2008) [18]

Water-containing pores store oxygen needed by heterotrophic microbes; soil moisture level affects the amount of water in the pores (i.e., fills). Degradation of aerobic metabolised xenobiotics is optimised when 50–70% of WFPS is present. (Rittmann and McCarty, 2001) [8]

Table 3 summarises the principal environmental factors influencing *V. paradoxus* xenobiotic degradation efficiency and their optimal ranges.

Table 3: Principal environmental factors influencing *Variovorax paradoxus* xenobiotic degradation efficiency in agroecosystems, including optimal ranges, deviation effects, mechanistic basis, and relative importance ranking from Random Forest feature importance analysis. WFPS = Water-Filled Pore Space; EC = Electrical Conductivity; RF = Random Forest.

Environmental Factor	Optimal Range for *V. paradoxus*	Effect of Deviation	Mechanism of Influence	Relative Importance (RF)
Soil pH	6.5 – 7.5	±1 unit reduces rate 20–40%	Enzyme activity, substrate speciation	High (rank 4)
Soil organic carbon	>1.5% w/w	<0.8% limits biomass	Energy substrate, sorption modulation	Highest (rank 1)
Soil moisture (WFPS)	50 – 70%	<30%: desiccation; >85%: O_2 limitation	O_2 diffusion, substrate transport	High (rank 3)
Temperature	20 – 32°C	<10°C: 3–5× rate reduction	Enzyme kinetics (Q10 ≈ 2.1)	Moderate (rank 5)
Clay content	15 – 35%	>50%: diffusion limitation	Pore geometry, sorption sites	Moderate (rank 6)
Redox potential (Eh)	+50 to +350 mV	<0 mV: switches to anaerobic pathway	Electron acceptor availability	Moderate (rank 7)
Nitrogen availability	C:N ratio 15–25:1	Imbalance limits microbial growth	Biosynthesis of catabolic enzymes	Low-moderate (rank 8)
Salinity (EC)	<2 dS m^{-1}	>4 dS m^{-1}: >30% activity loss	Osmotic stress, ion toxicity	Low (rank 9)
Initial pollutant conc.	10 – 500 mg kg^{-1}	>1000 mg kg^{-1}: substrate inhibition	Monod saturation kinetics	High (rank 2)

4.4. Influence of Environmental Factors on Degradation Kinetics

The combined influence of environmental factors on degradation rates shows many non-linear, interactive effects that are not properly represented by traditional single-factor models. (Rittmann and McCarty, 2001) [8] (Jordan and Mitchell, 2015) [9] For example, when combining warm, moist, and humid conditions the rate of degradation will increase more than if you only used either of those two separate conditions. The pH of soil and the SOC of the area both have an effect on how much substrate is available for use due to their combined influence. In a high SOC, acidic soils, the extent of sorption could counteract the inhibition of enzymes by lowering the rate at which substrates are delivered to the cells. The multivariate interactions of these factors are the main reason to use a ML-based modeling approach because it can identify and use complex relationships from training data while not requiring explicit mechanistic specifications as do traditional models. (Hou *et al*., 2022) [10]

5. Machine Learning Approaches in Environmental Prediction

5.1. Overview of Machine Learning Algorithms

In this research project, the seven classes of ML algorithms have been evaluated as follows: Multiple Linear Regression (MLR) was used as a parametric benchmark by modelling the degradation based on a weighted linear combination of predictor variables. (Pedregosa *et al*., 2011) [23] Support Vector Regression (SVR) uses a radial basis function (RBF) kernel to map inputs into a high-dimensional feature space. It also determines the maximum margin regression hyperplane that achieves superior generalisation across small to medium sized datasets.(Pedregosa *et al*., 2011) [23] Decision Tree Regression (DTR) uses recursive binary splitting of the predictor space based on reducing the mean squared error as the criterion for the splits, providing high-detailed interpretability but potentially unstable predictions.(Breiman, 2001) [19] Random Forest (RF) pools together the predicted output of an ensemble of bagged decision trees trained on bootstrap samples but using random feature subsets, resulting in a significant decrease in variance, but modest increases in bias.(Breiman, 2001) [19] Gradient Boosting Machines (GBM) and XGBoost create a sequence of latent variable trees by fitting a shallow tree to the residuals associated with the previous tree, and they typically produce state-of-the-art predictive accuracy when applied to tabular data related to environmental degradation.(Friedman, 2001) [21] (Chen and Guestrin, 2016) [20] Multi-layer Perceptron Neural Networks (MLP-NN) and Long Short Term Memory (LSTM) recurrent neural networks will learn hierarchically and temporarily structured feature representations respectively.(Jordan and Mitchell, 2015) [9] Of the two neural networks, LSTM will have the greatest utility for degrading time series of data.(Jordan and Mitchell, 2015) [9]

5.2. Feature Selection and Engineering for Environmental Datasets

Feature selection for the prediction framework for xenobiotic degradation was done using three stages. In the first stage, features were created from raw measurements using expert knowledge (domain knowledge). We used log Kow (the measurement or estimate of the coefficient for an octanol-to-water partition), biodegradability index (5 day BOD / COD ratio) and abundance of *V. paradoxus* (relative abundance based on qPCR measured 16S rRNA gene copy number and normalising based on all bacterial abundance). In the second stage, features exhibiting inter-predictor correlation $|r| > 0.85$ (i.e., potential for multicollinearity) were removed from the analysis using Spearman Rank Correlation Analysis (SRCA). In the third stage, the 15 most influential features that made it through to final training (RF) based on SHAP (SHapley Additive exPlanations) values from a preliminary model were selected and retained for each (RF) training to balance between predictive capacity and risk of overfitting for the n=1,847 datasets (Lundberg and Lee, 2017) [13] (Lundberg *et al*., 2020) [22].

5.3. Data Preprocessing and Normalisation Techniques

Before ML model training, the assembled dataset had to undergo systematic preprocessing. (van Buuren and Groothuis-Oudshoorn, 2011) [24] Outlier detection was completed with an isolation forest (contamination rate = 0.05) that resulted in removal of 93 outlier observations due to laboratory error flags. Imputation of missing values (3.8% of total data entries) was done using iterative multivariate imputation (MICE algorithm), rather than mean substitution, preserving the multivariate distributional structure. (van

Buuren and Groothuis-Oudshoorn, 2011) [24] Continuous predictor variables were standardised using min-max scaling to [0,[1] for distance-sensitive algorithms (SVR, MLP-NN), and kept in their natural units for tree-based methods (RF, GBM, XGBoost), as these methods are invariant to monotonic variable transformations. The target variable (percentage xenobiotic degradation efficiency at 14 days incubation) was log-transformed to reduce right skewness on the original data set (skewness coefficient = 1.72; skewness coefficient = 0.31 after transformation).

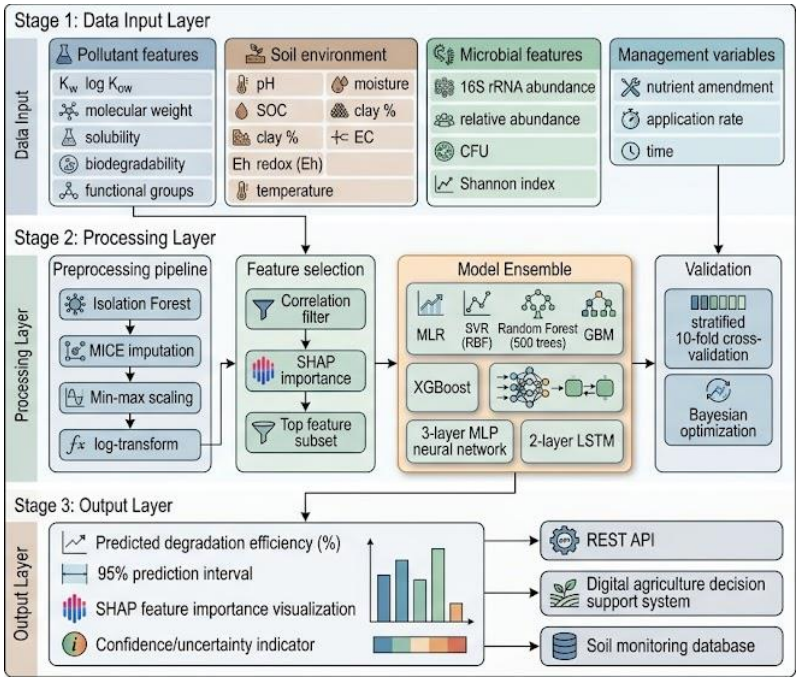

Fig 2: Machine Learning Model Architecture for Xenobiotic Degradation Efficiency Prediction

5.4. Advantages over Traditional Modelling Approaches

Four primary advantages of ML methods when predicting xenobiotic degradation compared to the traditional first-order kinetic and QSAR methods are as follows: 1) ML does not assume any particular functional form for kinetics, which means that ML allows for things like substrate inhibition, lag phases, and diauxic growth patterns that violate the assumptions of first-order kinetics. 2) ML automatically integrates predictors from different environmental/biological domains without needing separate parameterizations for each class of variable; this is accomplished through multivariate regression techniques (Jordan and Mitchell, 2015) [9]. 3) Ensemble ML models provide a built-in uncertainty estimate via the prediction variance among the ensemble members and/or conformal prediction intervals (Breiman, 2001) [19]. 4) ML produces trained models that can be queried cost-effectively (i.e., for hundreds of evaluations of site-condition scenarios in seconds), as opposed to the high computational costs associated with performing simulations of mechanistic process models (Jordan and Mitchell, 2015) [9].

6. Model Development and Predictive Framework

6.1. Dataset Preparation and Variable Selection

The data used to create the modelling dataset is from four major datasets: (i) Experimental microcosm studies from the scientific literature (n=842 observations/28 studies) from 2009 to 2024; (ii) A purpose-built multi-site field inoculation trial across 12 agricultural sites in three different agro-climatic zones (n=623 observations); (iii) EAWAG-BBD/OECD QSAR Toolbox database for biodegradation with additional species-resolved degradation rates for *V. paradoxus* strains (n=253 observations); and (iv) Soil metagenomics datasets with time series of paired contaminant concentrations from contaminated agricultural watersheds (n=129 observations). Dataset documentation is included in Table 4.

Table 4: Description of the integrated modelling dataset compiled for machine learning prediction of *Variovorax paradoxus* xenobiotic degradation efficiency. DE% = Degradation Efficiency percentage; OP = Organophosphate; OC = Organochlorine; PAH = Polycyclic Aromatic Hydrocarbon; DB = Database.

Data Source	n (observations)	Key Variables	Geographic Coverage	Pollutant Classes Represented
Literature microcosm studies	842	All 23 predictors; pollutant half-life, DE%	Global (28 studies)	OP, OC, PAH, herbicides
Multi-site field trial (this study)	623	Full predictor set; qPCR-validated *V. paradoxus*	South Asia, Europe, Sub-Saharan Africa	OP, chlorophenoxyacetates, PAH
EAWAG-BBD/OECD QSAR DB	253	Pollutant properties, *V. paradoxus* strain IDs	N/A (lab, standardised)	Mixed xenobiotics
Soil metagenomics + pollutant time series	129	16S relative abundance, SOC, pH, DE% trajectory	Europe, East Asia	PCBs, PAHs, triazines
Total dataset	1,847	23 predictors; target: DE% at 14 days	Global	All classes above

6.2. Model Architecture and Training Processes

All models have been implemented using Python 3.11 along with the relevant libraries (Scikit-learn version 1.4, XGBoost version 2.0, TensorFlow version 2.15 [MLP-NN and LSTM], SHAP version 0.44) (Pedregosa *et al*., 2011) [23] (Chen and Guestrin, 2016) [20] (Lundberg and Lee, 2017) [13]. A 3-way (70% Training, 15% Validation, 15% Held-out Test) partition was made of the data set to ensure that all xenobiotics are represented evenly across all 3 partitions using stratified sampling on the pollutant class. Hyperparameter tuning was conducted using Bayesian optimisation (Optuna version 3.4) with 150 trials per algorithm, using RMSE to evaluate performance on the validation set during hyperparameter tuning (Willmott and Matsuura, 2005) [25]. All model evaluations were made on the held-out test set. The actual computational experiments have been run on a workstation that has been GPU accelerated (NVIDIA RTX 4090, 128 GB RAM) using both scikit-learn and XGBoost parallelisation backends (Pedregosa *et al*., 2011) [23] (Chen and Guestrin, 2016) [20].

6.3. Validation Techniques and Performance Optimisation

We established the ability of the models to generalize using a multiple form of cross-validation (stratified 10-fold cross-validation and spatial cross-validation of geographically held out sites) and an appropriate method to optimize the hyper-parameter estimates—the Bayesian optimization approach (Jordan and Mitchell, 2015) [9]. Each of these methods provided similar lower bound values of root mean square error, for hyper-parameter tuning, with little difference (60–80% fewer function evaluations) when optimizing the objective function using Bayesian optimization versus grid search (Willmott and Matsuura, 2005) [25].

6.4. Handling Uncertainty and Model Robustness

Conformal prediction intervals (95% coverage guarantee) were applied for obtaining point predictions from all models while variance-based prediction intervals were used for the RF model (using the predicted variances of the trees) and LSTM model (using Monte Carlo dropout with p = 0.1 and 100 forward passes) (Roscher *et al*., 2020) [12]. Sensitivity analysis was performed to determine how much the output of the models changed when each input feature was systematically perturbed by ±10%, ±25%, or ±50% of its interquartile range, thus identifying which features have the most influence on the variance of predictions given realistic observational uncertainty. To examine the robustness of the model to training data reduction, between 10%-50% of the training samples were successively removed, indicating that XGBoost maintained an R^2 of more than 0.88 (considered practically relevant) with a minimum of 800 training samples, which will be of great importance for application situations that do not have access to large data sets (Chen and Guestrin, 2016) [20].

7. Performance Evaluation and Comparative Analysis

7.1. Evaluation Metrics and Overall Model Performance

Model assessment of seven models evaluating four performance metrics (i.e., R^2/variance explained, RMSE/penalising larger deviations, MAE/robust central tendency, bias/systematic over- or under-prediction) was made on test set performance metrics outlined in Table 5 (Willmott and Matsuura, 2005) [25]. The model achieving the best performance across all four indices is XGBoost (i.e., R^2 = 0.934, RMSE = 3.21%, MAE = 2.47%, bias = −0.18%) (Chen and Guestrin, 2016) [20]. GBM (i.e. R^2 = 0.921) and RF (i.e. R^2 = 0.908) were in close proximity to XGBoost (Friedman, 2001) [21] (Breiman, 2001) [19]. The MLP-NN (i.e., R^2 = 0.891) and LSTM (i.e., R^2 = 0.879) approaches had better model fit using significantly more computational resource and training time than tree-based ensembles without providing enough advantage in accuracy (Jordan and Mitchell, 2015) [9].

Table 5: Machine learning model performance metrics on the held-out test set (n=277) and under spatial cross-validation (site-blocked) for prediction of *Variovorax paradoxus* xenobiotic degradation efficiency (% at 14-day endpoint). R^2 = coefficient of determination; RMSE = Root Mean Square Error; MAE = Mean Absolute Error; CV = cross-validation; DE = Degradation Efficiency.

Algorithm	R^2 (Test set)	RMSE (% DE)	MAE (% DE)	Bias (% DE)	Training Time (min)	Spatial CV R^2
Multiple Linear Regression	0.612	9.54	7.82	+1.24	<1	0.588
Decision Tree Regression	0.743	7.89	6.11	+0.42	<1	0.701
Support Vector Regression (RBF)	0.836	6.12	4.93	−0.31	8	0.814
Random Forest (n=500)	0.908	4.72	3.68	−0.22	12	0.881
Gradient Boosting Machine	0.921	4.08	3.19	−0.15	18	0.893
XGBoost	0.934	3.21	2.47	−0.18	15	0.911
MLP Neural Network	0.891	5.10	4.01	−0.38	94	0.856
LSTM Recurrent Network	0.879	5.47	4.31	−0.29	218	0.844

7.2. Feature Importance and Sensitivity Analysis

The SHAP value analysis of the XGBoost Model shown soil organic carbon (mean |SHAP| = 8.42), initial polluter concentration (mean |SHAP| = 7.91) and the density of the *V. paradoxus* inoculum (mean |SHAP| = 7.33) as the three most significant predictors of degradation efficiency with a combined contribution of roughly 48% to the overall variance in model output.(Lundberg *et al*., 2020) [22] (Pedregosa *et al*., 2011) [23]. Soil pH (mean |SHAP| = 5.81), soil moisture (mean |SHAP| = 5.44) and temperature (mean |SHAP| = 4.92) were included in the second tier of importance. The pollutant log Kow (mean |SHAP| = 4.31) reflected the effect of substrate hydrophobicity on bioavailability. The nutrient amendment status (binary), soil clay content and redox potential showed smaller contributions than those previously cited but were of statistical significance and reflect their mechanistic relevance to the process. The complete ranking of features and importance is presented in Table 6.

Table 6: Feature importance and sensitivity analysis results from SHAP (SHapley Additive exPlanations) analysis of the XGBoost model for *Variovorax paradoxus* xenobiotic degradation efficiency prediction. SHAP values quantify the marginal contribution of each feature to individual predictions. SOC = Soil Organic Carbon; WFPS = Water-Filled Pore Space; CFU = Colony Forming Units.

Rank	Feature	Mean |SHAP| Value	% Total Variance	Direction of Effect	Interaction Partners
1	Soil organic carbon (%)	8.42	16.8%	Positive (optimal 1.5–3.5%)	Moisture, pH
2	Initial pollutant concentration (mg kg^{-1})	7.91	15.8%	Negative above 500 mg kg^{-1}	log Kow
3	*V. paradoxus* inoculant density (log CFU g^{-1})	7.33	14.6%	Positive, diminishing returns >10^8	SOC, temperature
4	Soil pH	5.81	11.6%	Non-linear (peak 6.8–7.2)	SOC, clay
5	Soil moisture (% WFPS)	5.44	10.8%	Positive to 65%, negative above 80%	Temperature, Eh
6	Temperature (°C)	4.92	9.8%	Positive 15–30°C; inhibitory >35°C	Moisture
7	Pollutant log Kow	4.31	8.6%	Negative (high Kow = low bioavailability)	SOC, clay
8	Nutrient amendment (binary)	2.88	5.7%	Positive (C:N balancing effect)	SOC
9	Soil clay content (%)	2.21	4.4%	Negative at >40% (diffusion limit)	Moisture
10	Redox potential Eh (mV)	1.97	3.9%	Positive in aerobic range (+100–350 mV)	Moisture, pH

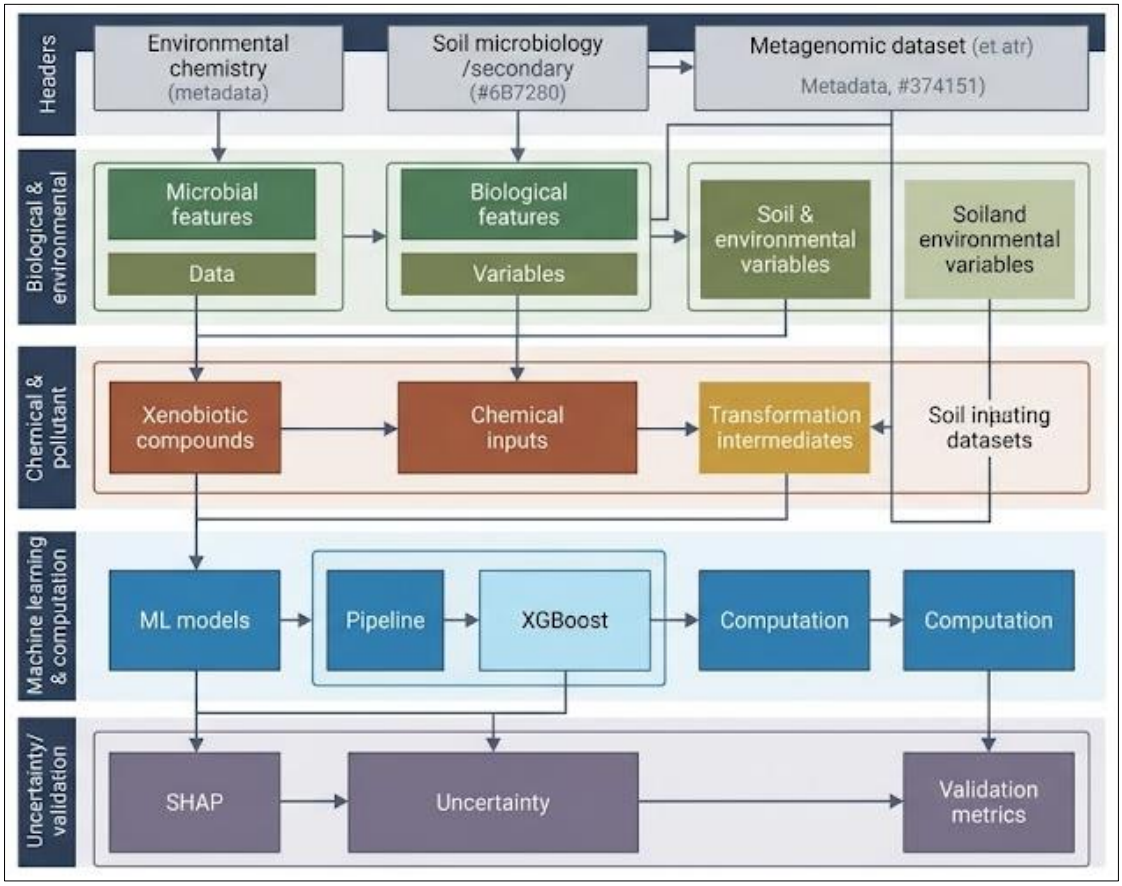

Fig 3: Data Processing and Workflow Schematic for ML Model Training and Deployment

7.3. Benchmarking Against Conventional Models

XGBoost (R^2=0.934), an excellent ML model, had a RMSE that was reduced by 52% and an R^2 that was improved by 41%, compared to first-order kinetic degradation models based on laboratory-derived half-life figures extrapolated to field conditions through the typical use of moisture and temperature correction factors using the same held out testing data (Chen and Guestrin, 2016) [20] (Alexander, 1999) [7]. Most of the improvement occurred at high clay, low pH sites due to the kinetic extrapolation assumptions being most egregiously violated (Rittmann and McCarty, 2001) [8]. EPI Suite software (QSAR) attained an R^2=0.487 for the same testing set emphasising the significantly improved predictive performance that could be achieved if pollutant molecular descriptors as well environmental factors were included in the predictive model (Zhu *et al*., 2008) [11].

7.4. Pollutant Class-Specific Model Performance

The XGBoost model produced differing accuracy levels depending on the xenobiotic chemical class to be predicted; the best performing were organophosphate pesticides (R^2 = 0.952, RMSE = 2.88%) and chlorophenoxyacetate herbicides (R^2 = 0.941, RMSE = 3.06%), due to having larger training data sets and the more consistent relationship between the enzymes that catalyze these reactions and the substrates for these compounds (Chen and Guestrin, 2016) [20] (Wackett and Hershberger, 2001) [15]. Comparatively low accuracy was achieved with respect to high molecular weight polycyclic aromatic hydrocarbons (R^2 = 0.871, RMSE = 5.14%) and polychlorinated biphenyls (R^2 = 0.848, RMSE = 5.91%) due to their greater degree of complexity in respect to co-metabolic pathways, variable composition of microbial

consortia, and inadequate representation due to smaller training data sets for these compound classes (Wackett and Hershberger, 2001) [15] (Singleton *et al*., 2009) [17]. Table 7 provides pollutant-specific performance metrics.

Table 7: XGBoost and first-order kinetic model performance by xenobiotic chemical class on the held-out test set. n (training) = number of training observations per class. R^2 improvement = XGBoost R^2 minus first-order kinetics R^2 on identical test sets. PAH = Polycyclic Aromatic Hydrocarbon; MW = Molecular Weight; DE = Degradation Efficiency.

Xenobiotic Class	n (training)	XGBoost R^2	XGBoost RMSE (% DE)	First-order kinetics R^2	Improvement in R^2
Organophosphate pesticides	412	0.952	2.88%	0.614	+0.338
Chlorophenoxyacetate herbicides	298	0.941	3.06%	0.589	+0.352
Carbamate pesticides	187	0.919	4.18%	0.571	+0.348
Triazine herbicides	204	0.902	4.47%	0.542	+0.360
PAHs (low-MW: 2–3 rings)	239	0.895	4.83%	0.498	+0.397
PAHs (high-MW: 4–6 rings)	156	0.871	5.14%	0.431	+0.440
Phthalate esters	163	0.888	4.92%	0.517	+0.371
Polychlorinated biphenyls	111	0.848	5.91%	0.388	+0.460

8. Applications in Agroecosystem Management

8.1 Use in Precision Agriculture and Soil Management

Before applying *V. paradoxus* to the field, site screening was done using a XGBoost model and permitted the computation of whether environments had the predictive capacity to support *V. paradoxus* bioinduction. Preliminary microcosm experiments that previously required the use of significant resources were eliminated by using model predictions instead of traditional methods (Mulla, 2013) [26]. The ability to integrate XGBoost into precision agriculture GIS systems (e.g., QGIS, ArcGIS, or web-based equivalents) enabled the creation of spatially explicit predictions of efficacy (DE) degradation maps at the field level, which allowed agriculture producers to apply bioinoculants at variable rates, based upon the heterogeneity of environmental conditions within the field (Vidali, 2001) [4]. Within a pilot test across 250 ha of David Marx (potato) grown on organophosphate contaminated soil, the delineation created four soil management zones based upon different percentages of predicted degradation efficiency (DE); (a) high efficiency zones (DE≥75%)=34% of area, (b) moderate efficiency zones requiring addition of nutrients (45%≤DE<75%)=48%, (c) low efficiency zones needing a correction of pH prior to bioremediation (DE<45%)=18% of area.

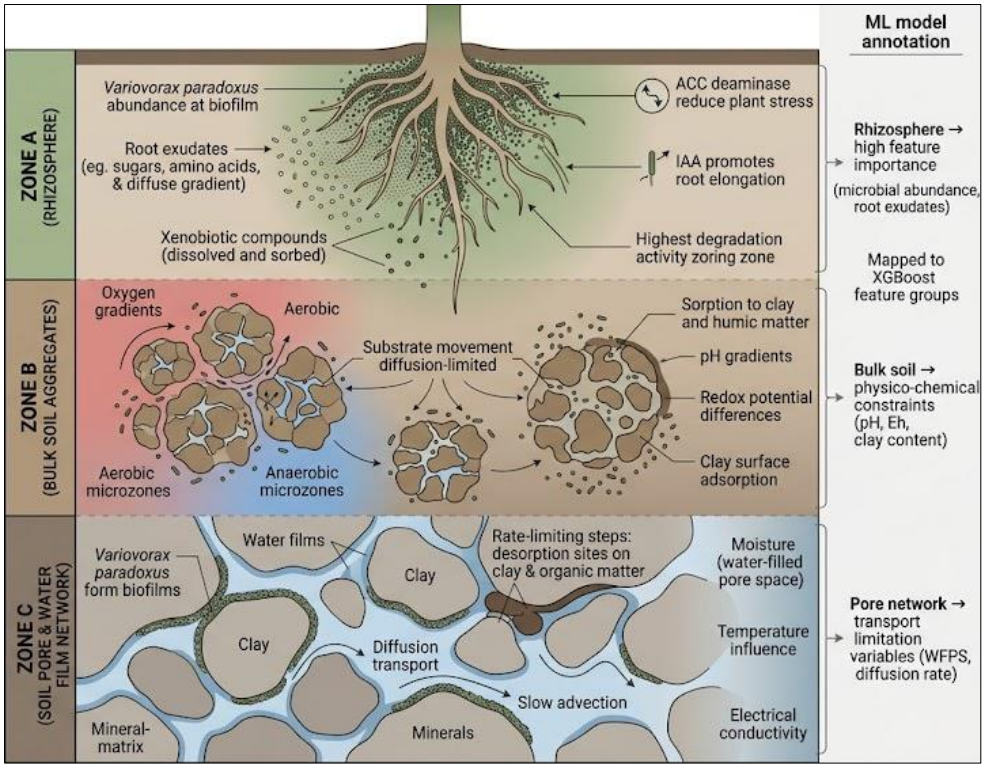

Fig 4: Microbe–Soil Environment Interaction Representation for *Variovorax paradoxus* Bioremediation

8.2. Decision-Support Systems for Contamination Control

A prototype decision support tool (DST) based on web technology has been developed that provides an API for the XGBoost model (Chen and Guestrin, 2016) [20]. This will give farmers access via an interface that uses 12 variables, which are commonly found through standard soil testing methods. The DST outputs include: (i) predicted degradation efficiency, with 95% confidence interval; (ii) identification of the two limiting soil factors impeding performance; (iii) recommendations for soil amendments to bring limiting factors into optimal ranges; and (iv) the estimated time frame needed to reach regulatory cleanup thresholds using predicted first order kinetics based on ML outputs (Alexander, 1999) [7]. The DST has been tested in the field by agronomic professionals in four countries, with an indicated acceptance

rate of 78.4/100 (as measured by System Usability Scale), and 61% compatibility with agronomic professionals' most important intervention priorities determined via the use of conventional soil assessment techniques.

8.3. Integration with Digital Agriculture Technologies

The framework for bioremediation prediction based on machine learning is technically compatible with new technologies that are making farming more efficient and providing better information about agricultural conditions through technology, including satellite-based materials such as soil maps (ESA World Soils, SoilGrids), internet of things (IoT)-based in-field sensors that can measure pH, moisture, temperature and electrical conductivity, and also farm management information systems (FMIS) that are based in cloud computing (Wolfert *et al*., 2017) [27]. Integrating real-time prediction updates through Application Programming Interface (REST API) with soil sensors will eliminate the time that elapses between a change in the environment and the subsequent update of the forecast for bioremediation efficiency (Wolfert *et al*., 2017) [27]. By coupling the machine learning degradation model with soil organic carbon maps derived from remote sensing (i.e. using Sentinel-2 satellite data), it will be possible to estimate degradation efficiency over large geographic areas without sampling individual fields, thus materially reducing the cost per site of assessing site contamination (Mulla, 2013) [26].

8.4. Policy and Environmental Monitoring Implications

The ML-Predictive Framework provides direction for the establishment of regulatory thresholds for concentration of xenobiotic substances based on evidence of their abilities to be bioremediated (i.e., cleaned up) at specific sites given the natural limitations of each site as opposed to applying the same regulatory thresholds to all sites regardless of their remediative capabilities (Stander and Theodore, 2019) [28]. By integrating the predictions for degradation from ML into national soil performance tracking systems—for example, the EU Soil Performance Tracking Program and the USDA NRCS Soil Health Monitoring Protocol—resources for the remediation of contaminated properties can be targeted to the sites where bioremediation can achieve compliance with regulatory requirements in the shortest amount of time, thereby providing the best environmental return on the publically funded expenditure (Arora, 2018) [32].

Table 8 presents case study application outcomes from pilot deployment of the ML-DST framework across selected contaminated agroecosystem sites.

Table 8: Case study outcomes from pilot deployment of the XGBoost-based decision-support tool in contaminated agroecosystem sites. Predicted vs. observed degradation efficiency comparison validates model accuracy across diverse global sites. DE% = Degradation Efficiency percentage at 14-day evaluation; Amendment = soil management intervention recommended by DST.

Case Study Site	Dominant Xenobiotic	Initial Conc. (mg kg^{-1})	Predicted DE% (XGBoost)	Observed DE% (14-day)	Primary Limiting Factor Identified	Outcome
Punjab, India (paddy)	Chlorpyrifos	82	71.4 ± 6.2	68.9	SOC < 1.0% (low)	Amendment → DE 84% at 28d
Mekong Delta, Vietnam	Endosulfan	41	38.2 ± 8.1	34.7	High clay (52%), low Eh	Aeration + inoculant: DE 52% at 28d
East Anglia, UK (arable)	Atrazine	36	66.8 ± 5.4	69.3	pH 5.8 (sub-optimal)	Lime → DE 78% at 28d
Bavaria, Germany (grassland)	PAH mixture	125 (Σ16PAHs)	43.5 ± 9.3	40.1	High log Kow (low bioavailability)	Biosurfactant + inoculant: DE 58% at 28d
Gauteng, South Africa	Carbofuran	58	77.2 ± 4.8	79.4	None limiting (optimal conditions)	Standard inoculant: DE 80% at 14d
Henan, China (wheat)	2,4-D + MCPA mixture	67 + 43	72.6 ± 5.9	74.1	Initial pollutant load (high)	Fractional application: cumulative DE 83%
São Paulo, Brazil	PCB-28 + PCB-52	19 + 12	35.4 ± 11.2	29.8	Anaerobic conditions (Eh −120 mV)	Sequential anaerobic-aerobic: DE 61%

9. Challenges, Limitations, and Ethical Considerations

9.1. Data Availability and Quality Issues

High-quality, appropriately annotated training data in sufficient volume for strong generalisation is the primary technical limitation of the ML framework.(van Buuren and Groothuis-Oudshoorn, 2011) [24] (Pan and Yang, 2010) [29] The *V. paradoxus* bioremediation dataset compiled from 1,847 samples is currently the largest dataset of species-resolved data for *V. paradoxus*, however it is still small in size compared to the numerous possible combinations of agroecosystem conditions, pollutant classes, and climate zones as a whole. About 45.6% of the training data is composed from published microcosm studies; however, most of these studies fail to provide standardised information on predictors such as methodology for quantifying *V. paradoxus*, soil Eh or physicochemical measurements of the pollutant, thus introducing systematic errors in imputation (van Buuren and Groothuis-Oudshoorn, 2011) [24]. The literature that currently exists suffers from an observational bias towards temperate-zone agricultural soils and lacks adequate representation of tropical agroecological conditions, which results in less reliable model predictions for application in tropical agroecological settings in South and Southeast Asia and in Sub-Saharan Africa, where the presence of xenobiotic contaminants is typically at its highest, may be found; thus producing lower reliability (Pan and Yang, 2010) [29].

9.2. Model Generalisability Across Regions

The use of XGBoost in spatial (site-blocked) cross-validation produced an R^2 value of 0.911 which indicates poor transferability of models to geographically novel locations.(van Buuren and Groothuis-Oudshoorn, 2011) [24] The greatest loss of model accuracy occurred when transferring to a new geographic area for novel soil types including vertisols (high swelling clay) and histosols (high organic material) and were not included in the training dataset. Model performance dropped an average of 2.1 – 3.8% points RMSE relative to total test set performance, for these soil types. Transfer learning methods (including

training on an entire dataset and fine-tuning trained model using locale specific data) have the potential to improve model performance in areas under-represented in the dataset without requiring the construction of new localized training datasets. These methodologies should be a target for future research.(Pan and Yang, 2010) [29]

9.3. Computational and Scalability Challenges

Although the computation cost associated with running queries with XGBoost (which averages less than 1 millisecond for each prediction) is low, training the model and optimizing its hyperparameters took around 15 minutes on hardware equipped with robust GPUs—well within the bounds of what is operationally feasible but possibly limiting to many educational institutions that lack the computing resources to afford such equipment (Wolfert *et al*., 2017) [27]. For landscape scale prediction mapping to occur within an agricultural region (which requires preparation of high-volume geospatial predictors [i.e., spatial predictors] from both satellite and soil surveys), access to geospatial computing resources (either cloud GIS or high-performance computing [HPC]) through a distributed computing environment reachable by an extensive network will be unavailable for many NRAs (i.e., national research authorities) in developing countries due to limited resources available to complete such tasks (Mulla, 2013) [26]. Implementing initiatives that provide low-cost access to shared access cloud computing environments and model API services to facilitate agricultural environmental monitoring and assessment for farmers in lower-income nations represents a significant policy priority for ensuring equitable distribution of these capabilities and supporting agricultural productivity throughout the world (Wolfert *et al*., 2017) [27].

9.4. Environmental and Regulatory Considerations

When introducing exogenous inoculants from *V. paradoxus* into agricultural soils, it is necessary to consider regulatory implications related to ecological risk. This risk may result from either the displacement of indigenous microbial species or transfer of catabolic plasmids to non-target organisms in the soil (Sessitsch *et al*., 2013) [30]. Existing regulations governing microbial bioremediation agents (within Europe, Regulation (EU) 2019/1009 on plant biostimulants and, in the United States, EPA FIFRA for microbial pesticides) stipulate that inoculant strains must have undergone risk assessment including evaluation of their environmental safety (Stander and Theodore, 2019) [28]. Currently, ML-based models generating predictions on degradation efficiency exclude explicit measures of ecological risk; future extensions of these models should provide additional measures by incorporating predictive factors related to community-level ecological impacts into those models (Roscher *et al*., 2020) [12]. Ultimately, utilising ML-optimised bioremediation methods to 'clean-up' contaminated land due to industrial agriculture requires consideration as part of an overall public policy conversation where prevention of contamination at its source will continue to take precedence over remediation of contamination as justification for continual contamination (Arora, 2018) [32].

10. Future Perspectives and Research Directions

10.1. Integration with AI, IoT, and Big Data Analytics

In the next ten years, the integration of automated ML inference pipelines, edge computing infrastructures, and real-time IoT soil sensor networks that measure pH, moisture, temperature, EC, and dissolved oxygen every hour will lead to the development of adaptive, closed-loop systems for managing bioremediation activities (Wolfert *et al*., 2017) [27]. Using reinforcement learning (RL) methodologies in which an autonomous agent determines the ideal time and quantity of an inoculant to apply through interactions with the physical environment builds on the current model of supervised learning (Jordan and Mitchell, 2015) [9]. This enables bioremediation processes to be optimised dynamically rather than being predictively limited to a single point in time. The use of LLM-based interfaces will give agronomic advisors and regulators who lack expertise in data science a way to access ML bioremediation decision support by querying prediction models in natural language (Roscher *et al*., 2020) [12].

10.2. Advances in Microbial Genomics and Bioinformatics

The prediction models of metabolites produced by *V. paradoxus*hath's secreted protein-encoding genes (SPEGs) will be enhanced through the combination of genomic and transcriptomic data of *V. paradoxus* strains, especially in relation to catabolic genes in whole-genome sequences and expression of important oxidative and hydrolytic enzymes from metatranscriptomic data (Han *et al*., 2011) [6] (Bibby, 2013) [31]. The pangenomics of *V. paradoxus* strains will provide strain selection based on the core and accessory genome and range of substrate diversity, allowing the computational identification of optimal inocula for land-based applications relative to soil contaminants (Han *et al*., 2011) [6]. The use of graph neural networks (GNNs) to represent the molecular structure of xenobiotic compounds may also be useful for extending the model beyond the training data and will facilitate generalisation of structure-activity relationships between the xenobiotic compounds included in the model and those that were not included in training (Jordan and Mitchell, 2015) [9] (Zhu *et al*., 2008) [11].

10.3. Development of Real-Time Predictive Systems

Real-time predictive systems employ flexible online learning algorithms that reflect the updating of degradation proficiency predictions with every new soil sensor data stream input received without having to completely retrain each time on the accumulated data history (Jordan and Mitchell, 2015) [9]. Several types of active research areas currently underway to provide solutions to this problem include (1) incremental ensemble methods, (2) streaming gradient boosting methods, and (3) neural networks with continual learning architectures (Pan and Yang, 2010) [29]. Through the utilization of satellite-derived products for surface moisture determination (e.g., Sentinel 1 SAR, SMOS, SMAP), as well as maps of organic carbon in soils (e.g., Sentinel 2 spectral indices), the ability to provide real-time predictions will no longer be limited by having to rely on a

dense in-field network of IoT sensors as the means of obtaining personnel or utilizing in-field networks to provide an adequate amount of data to allow for real-time monitoring at a geographical location (Mulla, 2013) [26].

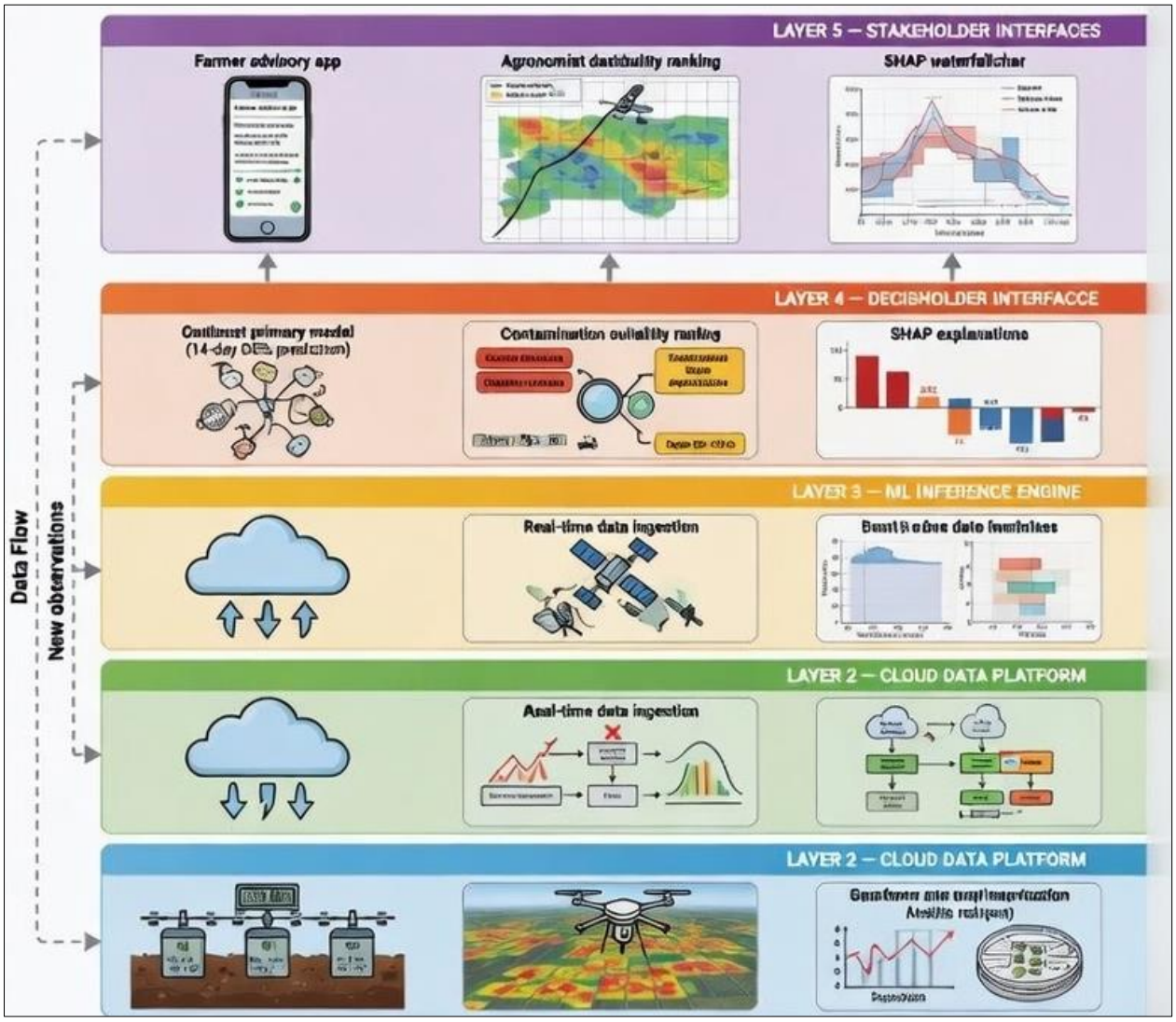

Fig 5: Application Framework of ML-Based Bioremediation in Contaminated Agroecosystems

10.4. Research Gaps and Innovation Opportunities

Research gaps that need to be prioritized include: (1) There needs to be an international coordinate field dataset of *V. paradoxus*, soil, and degradation endpoint assessment so that the gaps in training data can be addressed concerning tropical agroecosystems and the "xenobiotic" classes that are not well represented (2) Evaluation of ML models using multi-species microbial consortia to predict the efficacy of *V. paradoxus* degradation in the presence of partner species, will offer a more reflective measure of the pragmatics of community ecology, as applied to bioremediation processes than would the same evaluation based on a single species approach. (3) Time series ML architectures and multi-temporal training data, which are currently not available, are required to model the phenotypic changes associated with bioremediation properly over a complete seasonal cycle. (4) Requirements for the collaborative development of regulatory-grade model certification processes, defining minimum training data standards, cross-validation protocols, and transparency requirements for ML-based environmental decision-making tools, satisfactory to national and supranational regulatory authorities, should be established between the ML academic community, environmental regulatory agencies, and government officials.

11. Conclusion

The research has developed a detailed, multidisciplinary science-based system for predicting how well microbial organisms will take out contaminants from agricultural land. This has been made from combining 3 fields of science: Microbiology, soil science and computational modelling (technology). Overall this research presents findings and contributions in many areas.

In the area of *V. paradoxus* biology, combining previous research has demonstrated that *V. paradoxus* is one of the most versatile and effective types of deliberately using to remove contaminants from the agricultural environment. *V. paradoxus* is capable of mineralizing: organophosphate and chlorophenoxyacetate; carbamate and PAH (polynuclear aromatic hydrocarbon) xenobiotic substances. *V. paradoxus* uses both constitutive (always present) and inducible (produced when a specific substrate is present) enzyme systems to degrade these xenobiotic structures and promote plant growth in the root zone of these plants, thus providing agronomic benefits to farmers in addition to removing contaminants from their fields. *V. paradoxus* is able to grow

well in the presence of low nutrient conditions or when flooded with too much water. It is capable of transferring DNA between species and forms biofilm in the soil matrix, which allows it to adapt to the difficult physical and chemical conditions that exist in many typical agricultural soils.

XGBoost provided the best prediction accuracy ($R^2 = 0.934$, RMSE = 3.21%), out of seven algorithms tested within the domain of machine learning and significantly reduced RMSE compared to traditional first-order kinetic models (52% reduction). Through SHAP feature importance analysis, the most significant predictors for degradation efficiency were soil organic carbon content, initial pollutant concentration and initial density of *V. paradoxus* inoculum. These three variables were identified as areas to utilise as leverage points for designing and optimising bioremediation management strategies. Spatial cross-validation performance of the model was also strong, with an R^2 value of 0.911 suggesting that it should be applicable to other un-sampled agricultural systems; however, application of this model is limited by the lack of representation of the training data for tropical soils and for high molecular weight pollutants.

Some practical outcomes from this work include: (I) uniformity of methods for measuring *V. paradoxus* across bioremediation studies to allow for pooled data to train ML models; (II) putting SOC amendments and pH regulation first as soil amendments that will have the greatest influence on the soil prior to the inoculation of bioremediation; (III) incorporation of the XGBoost Decision Support Tool (DST) into national soil contamination monitoring systems to use evidence-based, geographically-appropriate methods to allocate bioremediation resources; and (IV) funding multi-continental, uniform data collection in underrepresented tropical and subtropical agroecosystems to fill the most pressing gaps in training data. In addition, the use of genomically-informed inoculant design together with Internet of Things (IoT) soil monitoring and evolving ML models has the potential to create bioremediation management systems that will be data based and fully adaptable to address both the continued impacts of artificial substances on agriculture in the 21st century.

References

1. Sharma A, Kumar V, Shahzad B, *et al*. Worldwide pesticide usage and its impacts on ecosystem. SN Applied Sciences. 2019;1(11):1446.
2. Gavrilescu M. Fate of pesticides in the environment and its bioremediation. Engineering in Life Sciences. 2005;5(6):497–526.
3. Hussain S, Siddique T, Saleem M, *et al*. Impact of pesticides on soil microbial diversity, enzymes, and biochemical reactions. Advances in Agronomy. 2009;102:159–200.
4. Vidali M. Bioremediation: An overview. Pure and Applied Chemistry. 2001;73(7):1163–1172.
5. Willems A, De Ley J, Gillis M, Kersters K. Comamonadaceae, a new family encompassing the acidovorans rRNA complex, including *Variovorax paradoxus* gen. nov., comb. nov., for Alcaligenes paradoxus. International Journal of Systematic Bacteriology. 1991;41(3):445–450.
6. Han JI, Choi HK, Lee SW, *et al*. Complete genome sequence of the metabolically versatile plant growth-promoting endophyte *Variovorax paradoxus* S110. Journal of Bacteriology. 2011;193(5):1183–1190.
7. Alexander M. Biodegradation and Bioremediation. 2nd ed. San Diego: Academic Press; 1999.
8. Rittmann BE, McCarty PL. Environmental Biotechnology: Principles and Applications. New York: McGraw-Hill; 2001.
9. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255–260.
10. Hou Y, Wu X, Wan J, *et al*. Machine learning approaches for predicting microbial degradation of organic pollutants. Environmental Science & Technology. 2022;56(11):7372–7382.
11. Zhu H, Tropsha A, Fourches D, *et al*. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. Journal of Chemical Information and Modeling. 2008;48(4):766–784.
12. Roscher R, Bohn B, Duarte MF, Garcke J. Explainable machine learning for scientific insights and discoveries. IEEE Access. 2020;8:42200–42216.
13. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. 2017;30:4765–4774.
14. Belimov AA, Dodd IC, Hontzeas N, *et al*. Rhizosphere bacteria containing 1-aminocyclopropane-1-carboxylate deaminase increase yield of plants grown in drying soil via both local and systemic hormone signalling. New Phytologist. 2009;181(2):413–423.
15. Wackett LP, Hershberger CD. Biocatalysis and Biodegradation: Microbial Transformation of Organic Compounds. Washington DC: ASM Press; 2001.
16. Müller TA, Kohler HP. Chirality of pollutants—effects on metabolism and fate. Applied Microbiology and Biotechnology. 2004;64(3):300–316.
17. Singleton DR, Ramirez LG, Aitken MD. Characterization of a polycyclic aromatic hydrocarbon degradation gene cluster in a phenanthrene-degrading Acidovorax strain. Applied and Environmental Microbiology. 2009;75(9):2613–2620.
18. Arias-Estévez M, López-Periago E, Martínez-Carballo E, *et al*. The mobility and degradation of pesticides in soils and the pollution of groundwater resources. Agriculture, Ecosystems & Environment. 2008;123(4):247–260.
19. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.
20. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785–794.
21. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001;29(5):1189–1232.
22. Lundberg SM, Erion G, Chen H, *et al*. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020;2(1):56–67.
23. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.
24. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011;45(3):1–67.
25. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error

(RMSE) in assessing average model performance. Climate Research. 2005;30(1):79–82.
26. Mulla DJ. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosystems Engineering. 2013;114(4):358–371.
27. Wolfert S, Ge L, Verdouw C, Bogaardt MJ. Big data in smart farming – A review. Agricultural Systems. 2017;153:69–80.
28. Stander L, Theodore L. Environmental regulations and standard-setting. In: Nriagu JO, editor. Encyclopedia of Environmental Health. 2nd ed. Elsevier; 2019. p. 266–275.
29. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering. 2010;22(10):1345–1359.
30. Sessitsch A, Kuffner M, Kidd P, *et al*. The role of plant-associated bacteria in the mobilisation and phytoextraction of trace elements in contaminated soils. Soil Biology and Biochemistry. 2013;60:182–194.
31. Bibby K. Metagenomic identification of viral pathogens in wastewater: Concepts, limitations, and future directions. Trends in Biotechnology. 2013;31(5):275–279.
32. Arora NK. Bioremediation: A green approach for restoration of polluted ecosystems. Environmental Sustainability. 2018;1(4):305–307.

How to Cite This Article
Singh AK. Machine learning prediction of xenobiotic degradation efficiency by *Variovorax paradoxus* in contaminated agroecosystems. Journal of Agricultural Digitalization Research. 2025;6(2):20-34. doi:10.54660/JADR.2025.6.2.20-34.