Review Article *Open Access*

Journal of Geoscience and Eco Agricultural Studies

ISSN: 3067-7297

DOI: *doi.org/10.63721/25JGEAS0113*

Machine Learning Algorithms for Soil Pollution Source Detection: A Systematic Review

Joseph Michael Odhiambo[1*] and Mvurya Mgala[2]

[1]Lukenya University, Kenya

[2]Technical University of Mombasa, Kenya

Citation: Joseph Michael Odhiambo, Mvurya Mgala (2025) Machine Learning Algorithms for Soil Pollution Source Detection: A Systematic Review. J.of Geo Eco Agr Studies 2(3): 01-20. WMJ/JGEAS-113

Abstract

Living organisms like plants, human beings, and microorganisms depend on pollution-free soil. Polluted soil has a great risk to the environment, especially agriculture, which provides livelihood to human beings. It also poses risks to the health of human beings all over the world. Tracking the source of soil contamination and determining the presence of contamination in soil best define environmental management and remediation. This systematic review identifies the application of machine learning (ML) algorithms in source identification and soil pollution categorization. Guided by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology, this study blends peer-reviewed papers published in the past ten years across various databases such as IEEE Xplore, ScienceDirect, SpringerLink, and Scopus. The review includes outstanding machine learning algorithms such as Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), k-Nearest Neighbors (k-NN), Neural Networks (NN), and ensemble methods, their efficacy, accuracy, required data, and interpretability. It also identifies the types of input data commonly employed (e.g., geospatial, physicochemical, remote sensing) and the most common feature engineering and model optimization methods. Trends suggest that there is a growing drift towards hybrid and deep learning approaches despite ongoing issues with model generalizability, data availability, and deployment in field conditions. The review is concluded by discussing current research gaps and suggesting future directions for robust, interpretable, and scalable ML-based soil pollution source detection systems.

***Corresponding author:** Joseph Michael Odhiambo, Lukenya University, Kenya.

Submitted: 03.09.2025 **Accepted:** 08.09.2025 **Published:** 18.09.2025

Keywords: Soil Pollution Sources, Machine Learning, Algorithm, AI, GIS

Introduction

Background and Significance of Pollution Source Detection

Soil health is a root component of environmental health, underpinning ecosystem resilience, agricultural productivity, and human well-being. Healthy soils feed plant growth, regulate water cycles, and provide a vital filter for contaminants [1,2]. Human activities have greatly compromised soil integrity by various means of pollution. The primary sources of soil pollution are industrial effluent disposal, agricultural sector runoff with chemicals and pesticides, the landfilling or dumping of waste in the open, and chemical or oil spills [3,4]. These contaminants bring with them toxic substances such as hydrocarbons, heavy metals, and persistent organic pollutants to infuse into the soil matrix, rendering long-term harm to biodiversity, food safety, and human health [5,6].

Identifying the specific sources of polluted soil is crucial to effective environmental cleanup and the establishment of meaningful policy responses [7,8]. Source attribution enables stakeholders to employ appropriate mitigation strategies, prosecute contaminators, and prevent repeat contamination events. Further analytical source detection of pollution assists in land-use planning and good resource management by allowing decision-makers to assess the source and spread of contaminants. Traditional soil sampling and laboratory analysis methods, as much as they are scientifically valid, possess some inherent shortcomings [9,10]. These traditional approaches are typically labor-consuming, time- consuming, and expensive, especially when applied to large-scale or distant regions. They typically provide discrete point data, which results in sparse spatial coverage and potential exclusion of minor-scale pollution gradients. Their passive nature further results in delayed action, thus making them less appropriate for dynamic and large-scale environmental monitoring. As a result, there is a need for quicker, more anticipatory, and scalable detection techniques of soil pollution sources. In this regard, new technologies such as machine learning offer the potential for such alternatives to break free from the limitations of traditional approaches to enable quick, less expensive, and broader analysis of the contamination pattern in the soil.

Role of Machine Learning in Environmental Monitoring

Machine learning (ML), which is one of the constituent areas of artificial intelligence (AI), is now a revolutionary approach to processing large and complex environmental data sets [11]. Using complex algorithms that learn from patterns in data, ML supports model building for outcome predictions, pattern discovery, and decision making without human intervention or with minimal direct human intervention [12]. These characteristics make ML highly suited to address environmental monitoring issues, wherein data is often big, heterogeneous, and dynamic. In the past several years, ML techniques have been utilized with tremendous success in a wide range of environmental applications. For instance, predictive models were developed for forecasting air quality levels in urban areas, anomaly detection in water quality parameters, and tracking changes in biodiversity from image and acoustic data analysis [13]. All these applications demonstrate that ML can process high-dimensional data efficiently, discover nonlinear patterns, and generate real-time insights that would be difficult to realize using traditional analytical techniques.

When applied to soil contamination, ML boasts a number of advantages over conventional methods. Unlike laboratory analysis and ground sampling, ML models are capable of merging data from heterogeneous sources, including remotely sensed imagery, geospatial data, and sensor networks, into source location models and pollution detection models [14]. This allows for an integrated comprehension of spatial and temporal pollution patterns. Again, the ability of ML to automatically perform repetitive data analysis and provide actionable findings in near real-time enhances environmental management process response and cost-effectiveness. Through the utilization of ML capabilities applied on source identification of soil contamination, researchers and practitioners can circumvent the limitations of traditional methods, enabling active, scalable, and data-intensive environmental surveillance systems.

Rationale for the Review

As soil pollution becomes increasingly prevalent with increased industrialization, urbanization, and unsustainable agriculture, the need for accurate, efficient, and scalable methods of pollution source detection is

growing. While traditional soil analysis techniques have provided valuable data, they are no longer sufficient to meet the needs of modern environmental management— especially when prompt, spatially extensive, and data-driven interventions are necessary. As a result, machine learning (ML) has become a desirable choice due to its promise of predictive modeling, automation, and pattern recognition in complex datasets [15].

While growing in quantity, research studies that apply ML to environmental applications are dispersed, and those specifically addressing source identification of soil contamination are dispersed [16]. These studies often vary considerably regarding the type of data, model approach, algorithm applied, and performance measures used. This variation in approaches renders comparison or generalization of opinions on best practices to apply ML methods in this application challenging. A systematic review should then be conducted to integrate current research activity, compare the performance of various ML algorithms, and define areas of ignorance that should be filled. The review here aims to synthesize the most relevant studies, criticize methodological approaches, and define the potential and limits of the use of ML for soil pollution source identification. In so doing, it presents a complete roadmap for researchers, practitioners, and policymakers wishing to take advantage of artificial intelligence in sustainable environmental remediation and monitoring.

Research Questions or Review Objectives

To comprehensively explore source identification of soil contamination with machine learning (ML), this current review is guided by a set of well-articulated and answerable research questions. They were created to define solid results, identify methodological patterns, and show literature gaps. The most obvious research questions guiding the current systematic review are:

- What are the most common machine learning algorithms used in source identification of soil contamination? This demands the most common forms of ML models employed within this application and reasons for such frequency.
- Most common form of data used with ML models in source identification of soil contamination and why so? Understanding of how input data nature—i.e., geospatial, physicochemical, or remote sensing—is and preprocessing thereof for ML modeling. Knowledge of the nature of input data and how they are preprocessed to fit into ML modeling will be required in order to ascertain model reliability and universality.
- How is the efficiency and effectiveness of these ML models established for use in source identification of soil pollution? This entails comparison of techniques used to establish accuracy, stability, and interpretability, and benchmarking techniques used in research.
- What are current trends, issues, and future direction of applying ML to detect the sources of soil pollution, with particular reference to interpretability of models, alignment with emerging technologies, and policy matters? After placing ML in context with regard to where it is being placed, i.e., how it stacks up with regard to actual need, existing policy, and future paths in data science and AI.

Collectively, the questions form the main objectives of the review, and in a combined and complete overview, how ML is transforming soil pollution observation and source apportionment.

Scope and Delimitations of the Review

This systematic review is specifically focused on machine learning (ML) algorithm applications towards source identification and soil contamination attribution. The review consists of peer- reviewed journal papers, conference papers, and technical reports published predominantly during the past decade in order to identify recent advancements and future directions of research in the area. The review contains experiments with supervised, unsupervised, and hybrid ML models with various environmental data sets ranging from geochemical, geospatial, remote sensing, and sensor data.

The attention of the review is specifically given to source identification rather than to measurement or estimation of overall soil quality or extent of pollution. While there may be some overlap, most emphasis has been placed upon those studies directly related to identification or classification of pollution sources-urban, agricultural, or industrial. Methodology, data preprocessing techniques, evaluation metrics, and

application of ML models to real-world scenarios are also addressed in the review.

But certain delimitations are noted. First, this review does not consider research based solely on traditional statistical or deterministic models without any machine learning. Second, unpublished theses, grey literature, and those published in languages other than English are not included, so regional diversity may be restricted by not considering them. Third, while review attempts to measure model utility and practitioner application, it is not able to conduct quantitative meta- analysis due to study design, data sets, and appraisal heterogeneity. These stated limits facilitate accurate and uniform analysis that allows for effective synthesis and interpretation of results for researchers, environmental practitioners, and policy makers.

Review Methodology

Review Protocol

This systematic review was conducted strictly according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, which provide a specific and systematic approach to identification, selection, evaluation, and synthesis of research studies [17]. Adherence to PRISMA provides methodological rigor, reduces bias, and enhances reproducibility in the review process. Protocol registration in an official registry such as PROSPERO was not performed, but all PRISMA principles provisions were literally followed at every phase [18].

Inclusion Criteria

We thought about including only studies for this review that met the following inclusion criteria:

- Relevance to Detection of Source of Soil Pollution: The study must be directly related to detection or soil pollution source classification, and not overall soil monitoring.
- Machine Learning Algorithm Application: The inclusion was limited to studies that applied supervised, unsupervised, semi-supervised, or reinforcement learning algorithms. They vary from the standard ML models such as Decision Trees and Support Vector Machines to sophisticated ones such as Deep Neural Networks and ensemble methods.
- Type of Pollution: Heavy metals (e.g., lead, cadmium, arsenic), organic pollutants (e.g., PAHs, VOCs), agrochemicals (e.g., herbicides, pesticides), hydrocarbons, and microplastics were the types of soil pollutants under consideration.
- Peer-Reviewed Journal Articles: Full-text peer-reviewed journal articles or academic conference proceedings only from credible sources.
- Publication Year Window: 2013 to 2024.
- Language: English language publications only.

Exclusion Criteria

The exclusion criteria applied are as follows:

- Inadequate Application of Machine Learning: Individuals using statistical, rule-based, or deterministic modeling alone (e.g., linear regression without the application of machine learning method) were excluded.
- Nonenvironmental Theme: Those articles with focus solely on non-polluting soils such as erosion, salinity, level of nutrients, or overall fertility were excluded unless as related to the determination of sources of pollution.
- Non-Primary Literature: Editorials, opinion pieces, technical comments, book chapters, and non-peer-reviewed grey literature were ruled out on academic quality.
- Methodological Opacity: Flawed research studies with no explicit description of the ML model, data employed, or evaluation process were ruled out to check replicability and quality.

The protocol offered a systematic and rigorous guideline to the process of selecting the best and highest-quality studies on machine learning application in soil pollution source identification. By employing well-defined inclusion and exclusion criteria, the review is transparent with methodological consistency ensured.

Search Strategy

To conduct a comprehensive and reproducible assessment of the relevant literature, a systematic search strategy was developed according to PRISMA guidelines. The aim was to identify high- quality peer-reviewed literature on the application of machine learning (ML) for source identification of soil pollution under various environmental conditions.

Databases Used

The academic databases utilized were as follows due to the fact that they provide extensive coverage of environmental science, engineering, data science, and interdisciplinarity:

- Scopus – Offers great coverage of scientific topics from environmental science to computer science.
- Web of Science – Renowned for its high-impact, peer-reviewed journals.
- IEEE Xplore – Technologically driven and specialized in nature when it comes to technology and engineering, great for machine learning and AI-based research.
- ScienceDirect (Elsevier) – One of the main sources of research in applied and environmental sciences.
- SpringerLink – Offers multidisciplinary journals with good emphasis on environmental modeling and AI.
- Google Scholar – Used to retrieve any other suitable or recently published material not included within the listed databases.

The databases were chosen between disciplinary strength (i.e., IEEE Xplore for ML, ScienceDirect for environmental studies) and cross-disciplinary coverage (i.e., Scopus, Web of Science).

Keyword Strategy and Boolean Operators

Three blocks of keywords constituted the search strategy:

- **Concept 1:** Soil Pollution/Contamination: (soil OR “soil quality” OR “soil health”) AND (pollut* OR contaminat* OR heavy-metal* OR pesticide* OR microplastic* OR chemical* OR toxic*)
- **Concept 2:** Source Detection/Identification: (“source detection” OR “source identification” OR “pollutant origin” OR “pollution tracking” OR “attribution” OR “fingerprinting” OR “hotspot detection”)
- **Concept 3:** Machine Learning: (“machine learning” OR “deep learning” OR “artificial intelligence” OR “AI” OR “neural network*” OR “support vector machine*” OR “random forest*” OR “ensemble learning” OR “clustering” OR “classification” OR “regression”)

These were then Booleaned together to form the final search string:

Final Search Query: ((soil OR "soil quality" OR "soil health") AND (pollut* OR contaminat* OR heavy-metal* OR pesticide* OR microplastic* OR chemical* OR toxic*)) AND ("source detection" OR "source identification" OR "pollutant origin" OR "pollution tracking" OR "attribution" OR "fingerprinting" OR "hotspot detection") AND ("machine learning" OR "deep learning" OR "artificial intelligence" OR "AI" OR "neural network*" OR "support vector machine*" OR "random forest*" OR "ensemble learning" OR "clustering" OR "classification" OR "regression")

Search Strategies Used

- Truncation (*) was utilized to capture a variation of word forms (e.g., pollut* would capture pollutant, pollution).
- Phrase searching (“\””) offered exact multi-word phrase matching like “machine learning” or “source identification.”
- Wildcards were not used due to database syntax variation but truncation allowed for term coverage.

Duplicate Management and Filtering

After record extraction from all of the databases, findings were exported to a reference management tool (i.e., Zotero or Mendeley) for deduplication. Duplication detection was performed based on title, authors’ names, and publication year and deleted automatically. Manual screening was performed for the second time with a view to eliminating false duplicates and ensuring unique inclusion of studies. Article titles and abstracts of the remaining articles were screened for exclusion and inclusion criteria (as outlined in Section 2.1), with additional full-text screening of included studies. A record of excluded studies and the reasons for exclusion was maintained to be transparent.

Selection Process

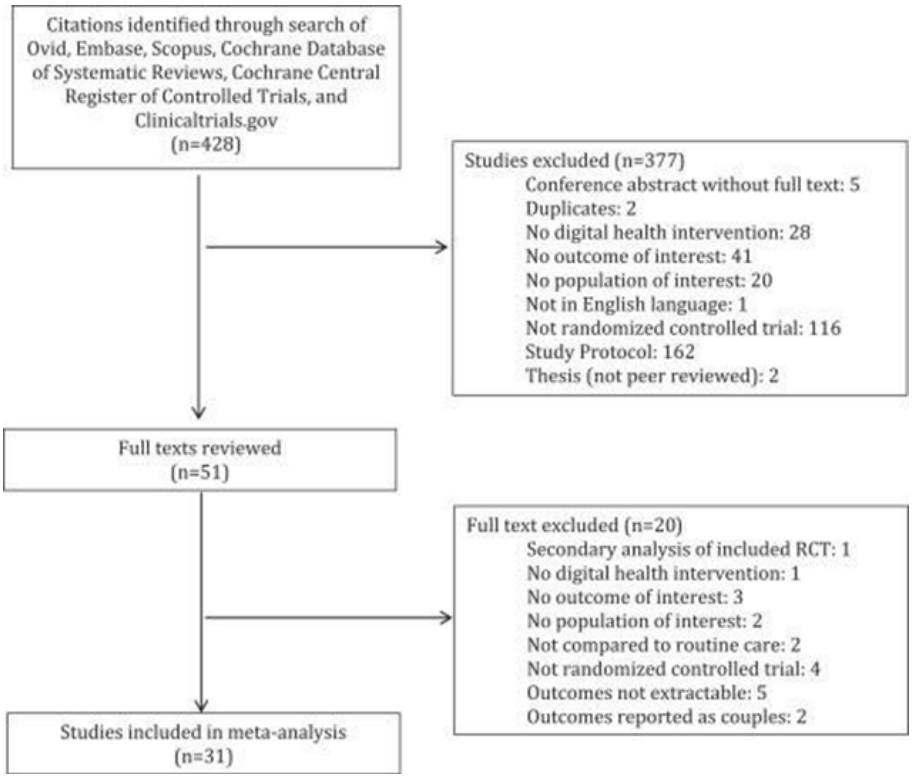
Study selection was subjected to a rigorous multi-stage screening to incorporate relevant and high-quality studies according to PRISMA guidelines. Database search and duplicate removal: Database search retrieved 1,600 records initially (Section 2.2). While importing records into a reference management system, duplicate records were found and removed, leaving a lesser number of 1,000 unique articles to be

screened further.

Title and Abstract Screening: Two reviewers conducted independent title and abstract screening for initial relevance to pre-specified inclusion and exclusion criteria (Section 2.1). At this point, those studies that were clearly out of scope of soil pollution source identification or lacked an ML focus were excluded. Where disagreement between reviewers existed, these were resolved through discussion to consensus. In this phase, 2,520 articles were excluded because of limited environmental relevance, non-ML method, or no data in abstracts.

Full-Text Screening: Full texts of the other 600 potentially qualified articles were independently downloaded and screened by the same two reviewers for eligibility based on the inclusion criteria. Articles that were not soil-specific, applied non-machine learning methods, or employed non- source detection as their goal were excluded. Exclusion reasons at this stage particularly were: "not soil-related," "no ML application," and "no source detection goal." 20 studies were excluded at this stage.

PRISMA Flow Diagram: To enable a simple graphical representation of the screening and selection process, a PRISMA flow diagram is constructed. It will illustrate the records retrieved, duplicates excluded, articles screened, full texts examined for eligibility, and studies ultimately included in the review. Publication of this flow diagram is a mandatory item of the PRISMA checklist, which adds transparency and reproducibility to the systematic review.



Quality Assessment

To determine the methodological quality and comparability of studies included in this systematic review, thorough quality appraisal was conducted. Considering that certain focus is being placed on machine learning (ML) techniques in environmental science, the quality assessment procedure was modified from available instruments such as the Critical Appraisal Skills Programme (CASP) checklist, Joanna Briggs Institute (JBI) tools, and the AMSTAR (A Measurement Tool to Assess Systematic Reviews) process. This was a mixed assessment targeted at significant points of concern to ML research, these being:

- Properly Framed Problem: Whether framed properly was the identification of soil pollution sources

and extent.

- Quality and Quantity of Dataset: Appropriateness and relevance of the dataset, e.g., representativeness, completeness, and dealing with missing values.
- Data Preprocessing and Feature Engineering: Appropriateness and readability of the data preprocessing methods, i.e., normalization, feature reduction, and feature selection.
- Explanation of why a particular machine learning approach is used under problem nature and data type.
- Performance Metrics and Validation: Associated evaluation metrics (e.g., precision, recall, F1-score, accuracy), and adequate validation methods like k-fold cross-validation or external validation.
- Reproducibility of Results: Adequate methodological information, code, or data to reproduce.
- Real-World Validation: Testing model performance against field data or an external database for the purpose of establishing real-world usability.

Evaluation Process

Two independent reviewers performed quality appraisal of all studies included to rule out bias and ensure consistency. In the event of disagreement on scores or interpretation, it was addressed through discussion or a third reviewer's referral. Quality scores directed the synthesis process by placing greater value on the strength of evidence from more robust studies in the narrative analysis. This intensive quality review process ensured that conclusions made from the review are based on solid and rigorous evidence and therefore provide more sound recommendations for future application and research.

Data Extraction and Synthesis

Data Extraction

A template extraction table was prepared with an effort to systematically extract the key information from each of the studies included. The following informations were extracted for enabling comparison and analysis in detail:

- Study ID, Author(s), Year of Publication
- Pollutant(s) Investigated: Type of soil pollutants that were investigated (e.g., heavy metals, organic pollutants, pesticides, microplastics)
- Type of Pollution Source: Categorization of the pollution sources such as industrial effluent, agricultural runoff, mining, or combined sources
- Location: Country or region in which the study was conducted
- Type(s) of Data Used: Categories of input data that were utilized for the modeling (for instance, sensor data, satellite data, lab soil data, GIS data)
- ML Algorithm(s) Used: Type of ML techniques employed (e.g., Random Forest, Support Vector Machine (SVM), Convolutional Neural Networks (CNN), k-Nearest Neighbors (k- NN))
- Problem Type: Type of ML problem (e.g., classification, regression, clustering)
- Key Features/Inputs: Most significant variables or features used for training
- Performance Measures Reported: Quantitative performance metrics such as accuracy, precision, recall, F1-score, Root Mean Square Error (RMSE), coefficient of determination (R^2)
- Main Results of Significance to Source Identification: Most significant results and conclusions drawn in relation to source identification of pollution
- Strengths and Weaknesses: Strengths and weaknesses as identified by the study
- Tools/Software Used: ML modeling environments or libraries (Python, R, TensorFlow, WEKA) employed

Synthesis Strategy

Because of anticipated heterogeneity in datasets, pollutant under study, ML algorithms, and performance measures anticipated, a thematic synthesis approach was employed. Data were synthesized into clusters thematically by category as follows:

- Machine learning algorithm types and comparative performance
- Data type differences and preprocessing strategies
- Pollutant types and respective detection challenges
- Study environmental and geographical conditions

The findings were subsequently described, compared, and interpreted by a narrative synthesis for these topics in an attempt to give a collective picture of state-of-the-art in ML-based soil pollution source identification

Quantitative meta-analysis in a formal sense was not feasible due to profound heterogeneity between studies. Heterogeneities in data sets, target pollutants, ML methods, and performance metrics meant that statistical pooling in a direct manner was not possible [19]. Narrative and thematic synthesis thus form an effective foundation to determine trends, areas of knowledge, and opportunities for research.

Results of the Review

Overview of Selected Studies

This chapter provides a descriptive summary of the studies selected for the systematic literature review by publication trends, type of pollutants investigated, and geographic location. This background information is used to provide a description of the current research direction and highlights how machine learning is being increasingly used to detect sources of soil contamination across the world. The chapter provides an introduction to the present research landscape, establishing primary trends in publication activity, polluting types, and geographic focus. These qualitative results set the stage for the comprehensive analysis that follows, particularly regarding how machine learning techniques are being utilized to inform source identification within soil pollution scholarship. This context of situation is critical to understand in terms of both the work that has already been accomplished and the problems that continue to persist with applying ML solutions under different environmental contexts.

Descriptive Statistics

100 studies were used and critically reviewed for this study. The studies span a wide period of publication, though it has significantly accelerated in the previous decade in terms of the number of publications. This boom is equated with the growing need for the application of artificial intelligence (AI) and machine learning (ML) techniques to environmental problems, e.g., pollution detection [20]. This coincides with the greater availability of open-source AI software, cheaper computation, and the availability of low-cost environmental sensors, all of which have made advanced modeling techniques within reach for everyone.

Literature that was assessed was made available in a very diverse range of academic journals, websites, and sites corresponding to the very interdisciplinary scope of this field of research. High- impact publications such as Environmental Pollution, Science of the Total Environment, Environmental Science and Technology, IEEE Transactions on Geoscience and Remote Sensing, and Journal of Environmental Management were cited as common publication outlets. This diversity showcases the intersection of a number of disciplines—environmental science, geospatial analysis, data science, and engineering—applied to identifying complex pollution issues through machine learning.

Pollutants Addressed

The majority of studies under review covered the identification and modeling of heavy metal pollutants such as lead (Pb), cadmium (Cd), arsenic (As), mercury (Hg), and chromium (Cr). These pollutants are typically found in industrial sites, mining operations, and agricultural fields impacted by fertilizers and pesticides. Subsequent research widened their scope to include organic contaminants such as pesticides, polycyclic aromatic hydrocarbons (PAHs), and volatile organic compounds (VOCs), causing more extensive damage to ecosystems and human health. Certain studies used a mixed-pollutant protocol and examined both inorganic as well as organic pollutants simultaneously to more accurately simulate real-world pollution environments.

Pollutant Type	Examples	Typical Sources	Study Focus
Heavy Metals	Lead (Pb), Cadmium (Cd), Arsenic (As), Mercury (Hg), Chromium (Cr) Pesticides, Polycyclic	Industrial sites, mining activities, agriculture (fertilizers, pesticides)	Identification and modeling of heavy metal contamination in soil
Organic Pollutants	Aromatic Hydrocarbons (PAHs), Volatile Organic Compounds (VOCs)	Agricultural runoff, industrial discharge, fossil fuel residues	Assessment of ecological and health risks from organic soil contaminants
Mixed Pollutants	Both inorganic (e.g., heavy metals) and organic (e.g., PAHs, VOCs)	Urban, peri-urban, and agro-industrial zones	Simulation of complex, real-world soil pollution scenarios for better accuracy

Table 3.1 summarizes the range of pollutants that were covered in the selected studies, focusing on the dominance of heavy metals and more recent issues such as microplastics—a relatively recent focus within soil pollution science. These newer contaminants also present additional modeling challenges due to their complex chemical behavior and detectability, offering additional scope for machine learning methodology to deploy [21].

Geographic Distribution and Environmental Context

The research canvassed was geographically spread over Asian, European, and North American research, with high concentrations of Chinese, Indian, United States, and Western European work. These are environmental "hotspots" either through focused industrial activity or ongoing environmental monitoring. Figure 3.2 maps out the spatial distribution of the investigated sites, demonstrating how study interest is globally widespread but locally focused on areas with higher pollution concerns.

While the primary focus of all the included studies was soil contamination, some extended their analytical reach to encompass multi-media environmental parameters. These extended studies incorporated measurements of air and water quality, as well as soil, to provide a more unified image of contamination pathways. This incorporation of the additional variables did not interfere with the core purpose of source apportionment of the soil contamination. Rather, it enhanced the analysis by illustrating the unification of environmental compartments.

Machine Learning Techniques Applied

The research at hand has the broad applicability of machine learning (ML) techniques to soil contaminant source identification with the predominant control by supervised learning techniques. Among them, Random Forest (RF) is the most commonly used technique [22]. RF is utilized because it has the ability to handle environmental noisy and high-dimensional data effectively and provide interpretable results in terms of ranks of feature importance [23]. Support Vector Machines (SVM) also prevail in literature for being particularly valued for their precision in classification issues as well as their capacity to cope with intricate decision boundaries even without large sets of data [24-26].

Artificial Neural Networks (ANN) are used widely because they can detect nonlinear relationships characteristic of environmental processes [27,28]. Deep learning techniques such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have been immensely used in the last couple of years, especially in studies that make use of spatial and temporal data [29]. CNNs provide optimal handling of spatially organized inputs like satellite pictures or geospatial maps, while LSTMs are able to effectively detect temporal patterns in time series data utilized for modeling pollution patterns [30,31].

Unsupervised learning methods, though less common, have been used in exploratory data analysis. Methods such as K-means clustering are used to cluster soil samples into similarities and identify areas where pollution has taken place or trends differ from what would be so in the absence of labeling [32,33]. Dimensionality reduction and feature extraction have utilized Principal Component Analysis (PCA) in an attempt to enhance the performance of models [34-36]. Ensemble methods that involve running several algorithms simultaneously, such as boosting and bagging methods, are common for improving prediction capability and generalizability [37,38]. Sequential model performance improvement using boosting algorithms such as Gradient Boosting Machines or XGBoost is employed for dealing with difficult-to-predict scenarios, while bagging methods such as Random Forest address variance through the aggregation of numerous decision tree predictions.

The algorithms are implemented in various types of machine learning issues in soil contamination research. Classification issues are the basis, where models are providing categorical sets back to sources of contamination, i.e., industrial, agricultural, or background natural [39]. Regression techniques, though with smaller demand, are used in estimating the concentration of contamination or predicting the relative contribution of various sources of contamination [40]. Cluster analysis allows for natural soil groupings of data to be meaningful in the identification of new patterns or hotspots irrespective of access to the source categories [41,42]. Algorithm choice has a direct correlation to data nature and problem to be solved. Random Forests and SVMs continue to be popular as they are reliable and simple to interpret in classification problems, whereas deep learning models progressively find application in modeling intricate spatial and temporal data [43-46]. This heterogeneity mirrors the increased complexity and sensitivity of machine learning methods in identifying the source of soil contamination.

Data Types and Sources

The reviewed literature used various types of data as inputs to machine learning algorithms for the determination of soil source pollution. The data sources in the majority of the cases are largely in- situ sensor measurement types, remote sensing images, GIS data, environmental history records, and, in some use cases, simulation data [47]. In-situ sensor measurements are largely used because they are first-hand measurements of the pollutants and soil parameters [48]. Electric conductivity, pH, and heavy metal content sensors provide instantaneous, localized measurements of worth in determining the magnitude of the pollution [49]. These measurements are the basis of the majority of ML models addressing point-based or small- area soil sampling.

Satellite and remote sensing have been increasingly important, especially for environmental monitoring of extended spatial areas [50]. Multispectral and hyperspectral sensors, mounted on Sentinel and Landsat satellites, sweep intense spectral information in varying wavelengths, and through indirect deduction, the existence of pollutants and soil types can be ascertained [51]. LiDAR measurements also supplement the spatial information by allowing topography mapping with accuracy, which helps in marking the pollutant dispersion patterns [52]. These data enable the identification of polluted locations in wide geographical areas outside of the spatial limitation of ground sampling. Geographic Information System (GIS) data facilitate analysis by the inclusion of contextual spatial covariates such as land use, topography, proximity to industrial facilities, road infrastructure, and water bodies [53,54]. These layers provide critical environmental and human context to augment the accuracy of source attribution models by closing the gap between observed pollution and probable human activities or natural conditions.

They also incorporate prior environmental information in the majority of studies. Some of such information includes prior history of monitoring, industrial permits, waste disposal reports, and other records of regulations that contain temporal information and ancillary information for pollution events and sources. Incorporation of historical data enables models to determine long- term trends of pollution as well as distinguish between recent and remaining contamination. Where partial or limited real-world data exist, other scientists have utilized simulated data to train and test models within controlled environments [55]. These simulated data sets enable testing for the ability of the ML algorithms to generalize and detect pollution under varying hypothetical scenarios.

Before applying such heterogeneous data to train machine learning models, strict preprocessing and feature engineering phases are needed. Original data tends to be normalized or standardized to normalize the scales, and missing data imputation techniques [56]. Outlier removal improves the data quality to avoid noise that can confuse the learning models [57].

Feature engineering is required to enhance model performance by creating novel, more informative features from available data [58-60]. For instance, spectral indices from satellite images such as the Normalized Difference Vegetation Index (NDVI) can be used as surrogates for soil health and pollution [61]. Spatial interpolation methods interpolate values at unsampled locations, providing added spatial information to point data. Temporal features extracted from time-series data retain the temporal dynamics of contaminants [62]. To manage high-dimensional information and prevent overfitting, research normally uses feature selection and dimension reduction strategies. One widely used approach is Principal Component Analysis (PCA) to reduce the number of variables but retain most of the variance. Other techniques like t-Distributed Stochastic Neighbor Embedding (t-SNE) and correlation-based filters aid in the selection of the most effective features, increasing computational effectiveness and model interpretability [63,64].

Performance Evaluation

Evaluating the performance of machine learning models in identifying the source of soil pollution is a range of measures contingent upon the nature of the problem—classification, regression, or clustering [65,66]. For classification problems, common in this research discipline, the most common measures are accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) [67]. Accuracy measures the global rate of correctly classified instances, and precision and recall measure about the model's ability to correctly report pollution sources with no misses or false alarms [68]. F1-score is the balance between precision and recall and provides a single score for imbalanced data sets. AUC-ROC measures the discriminative power of the model at varying levels of classification thresholds and is hence easy to compare binary or multi-class classifiers.

In applications where the problem has been framed as a regression task—i.e., pollutant concentration or estimation of source contribution—performance metrics have been founded on metrics like the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). Both RMSE and MAE measure the average magnitude of prediction errors, with RMSE placing larger errors at greater weight [69]. R^2 is the proportion of variance in observed data accounted for by the model and a goodness-of-fit measure. More random but used clustering analyses for pattern determination of soil contamination use indicators such as the Silhouette score and Davies-Bouldin index. These indicators assess the compactness and separation of clusters and confirm determined pollution clusters.

Performance comparison between studies shows that Random Forests usually perform better since they are robust to noisy and heterogeneous data, particularly in detecting heavy metals. Deep learning techniques such as CNNs are potential with spatially dense data such as satellite imagery but require additional data volume and computational resources. Support Vector Machines are suitable for small structured data but become overwhelmed by large spatial data. Ensemble methods and hybrid models tend to outperform individual algorithms through complementary strengths. Inter-study direct comparison is avoided by heterogeneity in datasets, types of pollutants, locations, and evaluation methods. Data heterogeneity in quality, preprocessing, and reporting of performance measures makes comparison challenging. Notwithstanding that, a few studies of within-study comparative tests, ensemble methods, and Random Forests were top performers overall [70].

Interpretability and Practical Use

In the past 10 years, the importance of Explainable Artificial Intelligence (XAI) has rapidly and amazingly, especially in environmental contexts, such as identification of sources that contribute to the soil pollution [71]. Even though machine learning systems are improving, especially with the emergence of deep learning, prediction and decision-making need to be clear and easy to understand, that is transparent. Such transparency is also critical to the establishment

of policymaker, environmental regulator, and other stakeholder trust, which is reliant on such models for informing remediation activities and regulatory decision-making [72].

There have been numerous XAI techniques employed throughout the reviewed studies above in an attempt to enhance explainability of the model [73]. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) allow researchers to explain a single prediction by attributing each input feature's contribution towards the model's output [74]. Tree-based predictors like Random Forests provide feature importance scores by default, which are the most significant soil or environmental variables that take part in source classification or prediction of the pollution level. Mechanisms of attention have also been used in deep learning for guiding the model to focus on the corresponding spatial or temporal regions from input data for better interpretability.

These interpretability methods aid in determining the most significant environmental elements that are causing pollution and where they might be coming from, thus closing the loop between informative knowledge and black-box models. For example, XAI can shed light on the way in which proximity to industrial sites or certain chemical soil characteristics influence pollution detection, which can then facilitate effective interventions.

Several case studies throughout the literature searched have shown applied influence through moving beyond model development in math to pilot testing or field use. One example is where ML models had been built into environmental monitoring systems, assisting regulators in finding areas of contamination or tracing sources of pollutants more efficiently than using current methods [75]. These analyses exemplify the real value in combining machine learning with interpretability methods in decision-making and environmental management enhancement. Incorporation of explainable AI method is crucial in transforming advances in machine learning into actionable soil pollution source identification tools usable by stakeholders and initiating proper actions in the environment.

Discussion

Trends and Advancements

The recent review of literature displays a few general directions and breakthroughs characteristic of the machine learning state of the art for source identification of soil pollution. One of the directions has been moving away from the classic machine learning algorithms such as Random Forests and Support Vector Machines towards more complex deep architectures. This has been enabled, in large measure, through advances in computing power and the increasing availability of huge, high-resolution environmental data to use for training advanced neural network models [76].

At the same time, remote sensing and Geographic Information System (GIS) data have been fused more into ML models. Fusing multispectral and hyperspectral satellite images, LiDAR images, and spatial context data enhanced the degree to map soil pollution in vast geographical areas with high precision and detail [77]. These technologies also promise constant surveillance of vast and remoteness-plagued areas, unencumbered by restrictions of ground sampling. Another basic advancement is the spatiotemporal models which deals with both spatial and temporal soil pollution dynamics. When time series data is utilized together with spatial characteristics, the models provide greater understanding of patterns in pollutant diffusion. They also help shifting locations of contamination hotspots to facilitate proactive environmental management.

Also, the application of techniques in increased demand manifests the emphasis in general towards transparency and interpretability [72]. While stakeholders demand greater responsibility from AI models, the creation and use of techniques like have become increasingly pertinent to ensure that deep models offer actionable, interpretable knowledge. These trends are characteristic of an advanced research area where increased data availability, computational power, and methodological improvement synergistically combine to enhance the performance and capabilities of machine learning in tracking the origins of soil pollution [74].

Gaps in Existing Literature

Despite some fairly major developments in the application of machine learning to source detection of soil

contamination, there remain some fairly major research gaps. Most notably is the absence of standardized datasets that can serve as benchmarks against which ML models can be tested and evaluated [78]. The heterogeneity of data source, pollutant types and geographical setting serves to limit replicability and generalizability of findings [79]. Some groups of pollutants, in particular complex mixtures that contain both organic and inorganic pollutants, continue to be poorly investigated. Most research is focused on individual classes of pollutants such as pesticides or heavy metals, which is reductionist given the multidimensional reality of real soil pollution. Another significant deficiency is minimal focus on identifying multiple co- occurring pollution sources. The majority of existing models are designed to classify or quantify a single source, but in fact soils are affected by overlapping pollution from co-occurring activities requiring more sophisticated multi-source attributional approaches [80].

The capacity of the ML models to generalize across the range of variable geological and environmental conditions is not dealt with satisfactorily as well. Models trained with data from one area perform poorly when applied in other areas due to variations in soil types, climate, and anthropogenic activities, indicating the necessity for adaptive or regional modeling approaches. The discipline also lacks long-term monitoring experiments that are instrumental in determining temporal trends in pollution dynamics and confirming predictive models over time. The developing world, where pollution is generally high and data gaps are a serious limitation, is also under-represented in studies. The geographical skew restricts the universal applicability of existing methods. The majority of studies emphasize predictive accuracy rather than explicit source attribution, and the inquiry is whether pollution is present or not, rather than untangling complicated sources and pathways of pollutants [81]. Closing this gap will entail combining domain knowledge and state-of-the-art ML methods to enhance interpretability and real-world usefulness.

Limitations in Datasets/Methodologies

Literature in consideration has a number of shared limitations on methods and datasets that discourage machine learning development for source detection of soil contamination. Key among the limitations are the quality and availability of data. Spatially broad, temporally resolved high- quality data are sparse. Most studies are localized or small-scale on datasets that don't capture variability in soil contamination and therefore limit model generalizability to large-scale use. Imbalance of data is also extremely critical. Dirty zones typically form a minority compared to clean or less-dirty zones, and therefore their databases get imbalanced, and consequently machine learning algorithms can become biased towards majority classes, and therefore lose their sensitivity to detect contamination accurately.

It is also difficult to obtain valid ground truth data to train and test models. Testing for soil contamination typically involves costly and time-consuming sampling and lab analysis as a constraint on confirmed contamination label availability. Unavailability undermines model prediction confidence and performance evaluation as well. Methodological differences introduce added complexity. There is considerable heterogeneity in method reporting approaches of studies, experimental setups, and performance metrics. It makes systematic synthesis of results as well as result comparison difficult. Heterogeneity is higher as differences vary among preprocessing steps, feature choice, and validation procedures. The majority of the prevalent machine learning models, particularly those based on deep learning, are not interpretable. These "black box" models provide minimal transparency into feature contributions to predictions, making it challenging to interpret, trust, and act on model outputs. This constraint underscores the importance of including explainable AI methods for enhancing transparency. Mitigation of these constraints is important in striving towards more stable, robust, and interpretable ML solutions for source attribution of soil pollution.

Emerging Technologies

Emerging machine learning and artificial intelligence technologies have great potential in overcoming the current challenges and innovating in the research field of soil pollution source identification. The new techniques offer distinctive capabilities that can be utilized to enhance data fusion, model accuracy, privacy protection, and adaptive environmental monitoring. Federated Learning is a novel approach to collaborative

model training. By enabling various organizations or geographical sites to collaboratively train machine learning models without revealing sensitive raw data, federated learning avoids insurmountable privacy and data sharing roadblocks. It has the most use in the soil pollution monitoring task where environmental agencies, research institutes, and industrial parties possess insightful yet sensitive datasets. Federated architectures thus can enable more use of the data and more generalizable models without sacrificing data ownership and confidentiality.

Geospatial AI combines machine learning with sophisticated spatial analysis to take advantage of spatial dependencies and relationships in environmental systems [82]. This has the potential to improve the precision of contamination mapping and source attribution via formal quantification of spatial autocorrelation and patterns. It is best done with improved precision in pollution hotspot detection and pollutant dispersal relative to landscape features such as topography, hydrology, and land use. Graph Neural Networks (GNNs) offer powerful means to represent complex, networked environmental processes. GNNs can encode pollution sources, transport pathways, and receptors as graph nodes and edges, preserving multi-dimensional relationships in soil pollution systems [83]. It is particularly useful for tracing the movement of pollutant through interactive soil, water, and air media in terms of more integrated source identification and impact assessment.

Reinforcement Learning can be used as well to create adaptive environmental monitoring strategies. Agents of reinforcement learning are able to learn optimum sampling policies and resource allocation policies and adaptively modify monitoring activities to prioritize areas of high risk or emerging pollution events to be more responsive and efficient. Combined, these emerging technologies provide new opportunities for scaling up the accuracy, scalability, and applicability of machine learning-based soil pollution source inference. Their inclusion in future studies and applications can be anticipated to improve beyond the limitations of the present and enable more informed environmental decision-making.

Integration with Policy and Real-time Monitoring Systems

The application of machine learning to source identification of soil contamination has significant practical implications for environmental policy and governance. By providing accurate and timely source identification of contamination, ML-based models can inform more targeted policy action and enhance regulatory compliance. Decision-makers can use the insights to prioritize cleanup in hotspots, allocate resources optimally, and design evidence-based policy that addresses the sources of contamination. In addition to retrospective analysis, the integration of ML models into real-time monitoring systems provides the potential for proactive environmental management. Real-time monitoring systems, through their networks of sensors and automated analysis of data, have the potential to enable early warning of pollution incidents and support prompt response measures. Such potential is especially crucial in preventing widespread contamination and minimizing ecological and public health impacts [84,85].

However, the translation of machine learning research into deployable systems and actionable policy faces many challenges. Regulatory approval often requires models to be able to offer transparency, reliability, and reproducibility—qualities that not all current ML approaches have. Additionally, implementing the necessary data infrastructure for maintaining data collection, processing, and storage can be expensive. There is also a need to build capacity within environmental agencies to be in a position to interpret model outputs and substantively integrate them into decision-making. Despite these challenges, the intersection of state-of-the-art ML technologies and environmental governance can revolutionize soil pollution management, interventions of which will be more targeted, timely, and effective.

Conclusion and Research Gaps

Summary of Key Findings

This systematic review proves that collective machine learning algorithms, for example, Random Forest here and highly advanced deep-learning models like Convolutional Neural Networks, are now omnipresent in the majority of soil contamination source identification applications. These algorithms are highly efficient in harnessing a vast array of data sources such

as remote sensing imagery, in-situ sensor measurements, GIS data, and historical data in order to detect complex spatial and temporal pollution patterns [83].

Performance metrics as documented across researches show overall high accuracy and stability. Model performance is still challenging to compare like-to-like owing to variability in methods, dataset changeability, and varying practice in assessment. Explainable AI (XAI) methodologies usage is quickly trending higher, elevating model explainability and interpretation of drivers of environmental pollution [73].

Although great leaps have been taken, there are still gaps-most significantly, the lack of harmonized data sets, lack of regular multi-source attribution, and under-representation of diverse global geographical areas. New technology across domains like federated learning, geospatial AI, and graph neural networks can potentially narrow some of the gaps and advance research and real- world application.

Explicit Research Gaps and Open Problems

With the discussion presented in Section 4.2, certain research needs and hot issues are emergent as being among the key lines of future machine learning for soil pollution source localization research:

- There is an urgent need for developing open, standardized, and representative benchmark databases with the objective of enabling fair evaluation, comparison, and reproducibility of ML models between and within research communities and environmental settings.
- Robust, interpretable machine learning models that can differentiate well and segregate well multiple, correlated sources of contamination are in dire need to model soil pollution real-world complexity.
- Long-term validation and generalizability tests should be the focus of future research, where the ML models must be ensured to be accurate and reliable under other geographical conditions, soil types, and time periods.
- The combination of mechanistic environmental models with data-driven machine learning methodology has the promise of resulting in better interpretability and prediction capability.

The solution to these issues will be necessary in order to build upon the scientific basis and application worthiness of machine learning methods in environmental pollution monitoring and regulation.

Implications for Future

As machine learning evolves and environmental monitoring hardware becomes more advanced, it's easy to picture studies on the identification of sources of soil pollution taking some very interesting new directions. First and foremost is the development of hybrid models—techniques that combine the scientific heft of conventional environmental models with the flexibility and learning ability of machine learning. Whereas traditional models assume processes that are chemically and physically well characterized, they become less effective when confronted with complicated real-world data. Machine learning, however, is very effective at discovering patterns and predicting fairly quickly from large sets of data. The two methods can complement each other—enhancing the precision of prediction and enhancing our knowledge of the way pollutants migrate through the ground.

Another important area of emphasis in the future is making environmental science machine learning models more interpretable. A lot of ML models today are essentially "black boxes"—they produce results, but it is not always clear how they got there. This is a bad thing if those results are going to be used to guide environmental policy or public health policy. Researchers have to build models that not just return the right answers but also yield explanations that researchers, decision- makers, and even communities can understand. More transparent, these models will be, the more trust it will generate and better, wiser decisions we will make regarding how we keep and take care of our soil.

Third is the exciting field of federated learning, where the promise is to enable various organizations to cooperate with each other without compromising sensitive information. In the majority of cases, high-quality environmental data are valuable in other places or entities that cannot exchange it due to privacy or ownership concerns. Federated learning allows such participants to develop robust models cooperatively—never exchanging raw data. Such a strategy not only helps maintain data privacy but also produces more generalized,

globally beneficial models that are coherent and compatible with different settings and geographies. These new developments—hybrid models, enhanced transparency, and shared use of data—can help push soil pollution research further. With the overcoming of current shortcomings and the embracing of these innovations, scientists will be better equipped to properly identify sources of contamination, facilitate effective cleanups, and make a meaningful contribution to environmental health and sustainability.

Proposed Direction for Research

In the future, the grand challenge in the next five to ten years for machine learning-driven source identification research of soil contamination is developing coupled, scalable, and understandable frameworks that bridge the divide between advanced technology and on-the-ground environmental issues. Making the above vision a reality is to foster interdisciplinary collaborations that integrate machine learning, environmental science, geospatial analysis, and policy scholars in codesigning both scientifically robust and socially effective approaches.

Future research must address real-world deployment of ML-based monitoring systems from conceptual models through pilot studies to in-place systems with the potential to deliver timely, actionable information to environmental managers and regulators. Integration of the systems into existing policy structures will be necessary to ensure effective conversion of technological advances into effective pollution prevention, remediation, and compliance enforcement.

This kind of realignment will require continuing investment in the development of standardized data infrastructure, open data sharing with privacy controls, and the creation of accessible tools for the benefit of stakeholders at all levels. By linking research to applied need and policy significance, the profession can make valuable contributions to soil health assurance, ecosystem conservation, and sustainable land use in an increasingly dynamic world.

References

1. Gavrilescu M (2021) Water, Soil, and Plants Interactions in a Threatened Environment. Water 13: 19 https://doi.org/10.3390/w13192746
2. Rashid A, Schutte B J, Ulery A, Deyholos M K, Sanogo S, et al. (2023) Heavy Metal Contamination in Agricultural Soil: Environmental Pollutants Affecting Crop Health. Agronomy 13: 6.
3. Hassan Al-Taai S H (2021) Soil Pollution—Causes and Effects. IOP Conference Series: Earth and Environmental Science 790: 012009.
4. Weldeslassie T, Naz H, Singh B, Oves M (2018) Chemical Contaminants for Soil, Air and Aquatic Ecosystem. In M. Oves, M. Zain Khan, & I. M.I. Ismail (Eds.), Modern Age Environmental Problems and their Remediation. Springer International Publishing 1-22.
5. Acharya S (2024) Heavy Metal Contamination in Food: Sources, Impact, and Remedy. In M. C. Ogwu, S. C. Izah, & N. R. Ntuli (Eds.), Food Safety and Quality in the Global South. Springer Nature 233-261.
6. Thakur S, Chandra A, Kumar V, Bharti S (2025) Environmental Pollutants: Endocrine Disruptors/ Pesticides/Reactive Dyes and Inorganic Toxic Compounds Metals, Radionuclides, and Metalloids and Their Impact on the Ecosystem. In P. Verma (Ed.), Biotechnology for Environmental Sustainability. Springer Nature 55-100.
7. Cachada A, Rocha-Santos T, Duarte A C (2018) Chapter 1 - Soil and Pollution: An Introduction to the Main Issues. In A. C. Duarte, A. Cachada, & T. Rocha-Santos (Eds.), Soil Pollution. Academic Press 1-28.
8. Khanam Z, Sultana F M, Mushtaq F (2023) Environmental Pollution Control Measures and Strategies: An Overview of Recent Developments. In F. Mushtaq, M. Farooq, A. B. Mukherjee, & M. Ghosh Nee Lala (Eds.), Geospatial Analytics for Environmental Pollution Modeling: Analysis, Control and Management . Springer Nature Switzerland 385-414.
9. Demattê J A M, Dotto A C, Bedin L G, Sayão V M, Souza A B e (2019) Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact. Geoderma 337: 111-121.
10. Upton R, David B, Gafner S, Glasl S (2020) Botanical ingredient identification and quality assessment: Strengths and limitations of analytical techniques. Phytochemistry Reviews 19: 1157-1177.
11. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta R K, et al. (2021) Artificial intelligence to deep learning: Machine intelligence approach for drug

discovery. Molecular Diversity 25: 1315-1360.
12. Sun A Y, Scanlon B R (2019) How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. Environmental Research Letters 14: 073001.
13. Olawade D B, Wada O Z, Ige A O, Egbewole B I, Olojo A, et al. (2024) Artificial intelligence in environmental monitoring: Advancements, challenges, and future directions. Hygiene and Environmental Health Advances 12: 100114.
14. Wang W, Wang G, Li J, Chen J, Gao Z, et al. (2025) Remote sensing identification and model-based prediction of harmful algal blooms in inland waters: Current insights and future perspectives. Water Research X 28: 100369.
15. Alotaibi E, Nassif N (2024) Artificial intelligence in environmental monitoring: In-depth analysis. Discover Artificial Intelligence 4: 84.
16. khatri A, kumar K, Thakur I S (2025) Emerging technologies for occurrence, fate, effect and remediation of organic contaminants in soil and sludge. Systems Microbiology and Biomanufacturing 5: 35-56.
17. Parums D V (2021) Editorial: Review Articles, Systematic Reviews, Meta-Analysis, and the Updated Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 Guidelines. Medical Science Monitor : International Medical Journal of Experimental and Clinical Research 27: e934475-1-e934475-3.
18. Booth A, Mitchell A S, Mott A, James S, Cockayne S, et al. (2020) An assessment of the extent to which the contents of PROSPERO records meet the systematic review protocol reporting items in PRISMA-P. F1000Research 9: 773.
19. Patial R, Sobti R C (2024) Exploring the Impact of Meta-Analysis in Scientific Research: A Review. Medinformatics. https://doi.org/10.47852/bonviewMEDIN42022447.
20. Chen H, Jia Q, Zhao X, Li L, Nie Y, et al. (2020) The occurrence of microplastics in water bodies in urban agglomerations: Impacts of drainage system overflow in wet weather, catchment land-uses, and environmental management practices. Water Research 183: 116073.
21. Fang S, Hua C, Yang J, Liu F, Wang L, et al. (2025) Combined pollution of soil by heavy metals, microplastics, and pesticides: Mechanisms and anthropogenic drivers. Journal of Hazardous Materials 485: 136812.
22. Salman H A, Kalakech A, Steiti A (2024) Random Forest Algorithm Overview. Babylonian Journal of Machine Learning 2024: 69-79.
23. Iranzad R, Liu X (2024) A review of random forest-based feature selection methods for data science education and applications. International Journal of Data Science and Analytics https://doi.org/10.1007/s41060-024-00509-w.
24. Manikandan G, Pragadeesh B, Manojkumar V, Karthikeyan A L, Manikandan,R, et al. (2024) Classification models combined with Boruta feature selection for heart disease prediction. Informatics in Medicine Unlocked 44: 101442.
25. Pisner D A, Schnyer D M (2020) Chapter 6—Support vector machine. In A. Mechelli & S. Vieira (Eds.), Machine Learning. Academic Press 101-121.
26. Singla M, Ghosh D, Shukla K K (2020) A survey of robust optimization based machine learning with special reference to support vector machines. International Journal of Machine Learning and Cybernetics 11: 1359-1385.
27. Chen Y, Song L, Liu Y, Yang L, Li D (2020) A Review of the Artificial Neural Network Models for Water Quality Prediction. Applied Sciences 10: 17.
28. Han K, Wang Y (2021) A review of artificial neural network techniques for environmental issues prediction. Journal of Thermal Analysis and Calorimetry 145: 2191-2207.
29. Wang S, Cao J, Yu P S (2022) Deep Learning for Spatio-Temporal Data Mining: A Survey. IEEE Transactions on Knowledge and Data Engineering 34: 3681-3700.
30. Tsokov S, Lazarova M, Aleksieva-Petrova A (2022) A Hybrid Spatiotemporal Deep Model Based on CNN and LSTM for Air Pollution Prediction. Sustainability 14: 9.
31. Yan R, Liao J, Yang J, Sun W, Nong M, et al. (2021) Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Systems with Applications 169: 114513.
32. Khorshidi N, Parsa M, Lentz D R, Sobhanverdi J (2021) Identification of heavy metal pollution sources and its associated risk assessment in an

industrial town using the K-means clustering technique. Applied Geochemistry 135: 105113.
33. Xu H, Croot P, Zhang C (2021) Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis. Environment International 151: 106456.
34. Hasan B M S, Abdulazeez A M (2021) A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. Journal of Soft Computing and Data Mining 2: 1.
35. Meng Y, Qasem S N, Shokri M, S S (2020).Dimension Reduction of Machine Learning-Based Forecasting Models Employing Principal Component Analysis. Mathematics 8: 8.
36. Parhizkar T, Rafieipour E, Parhizkar A (2021) Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction. Journal of Cleaner Production 279: 123866.
37. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects | IEEE Journals & Magazine | IEEE Xplore. (n.d.). Retrieved July 16, 2025, from https://ieeexplore.ieee.org/abstract/document/9893798
38. Rane N, Choudhary S P, Rane J (2024) Ensemble deep learning and machine learning: Applications, opportunities, challenges, and future directions. Studies in Medical and Health Sciences 1: 2.
39. Talukdar P, Kumar B, Kulkarni V V (2023) A review of water quality models and monitoring methods for capabilities of pollutant source identification, classification, and transport simulation. Reviews in Environmental Science and Bio/Technology 22: 653-677.
40. Liu G, Zhou X, Li Q, Shi Y, Guo G, et al. (2020) Spatial distribution prediction of soil As in a large-scale arsenic slag contaminated site based on an integrated model and multi-source environmental data. Environmental Pollution 267: 115631.
41. Gulledmath S, Hemanth K S (2024) Exploring Soil Diversity and Land Use Patterns in Arid Tropical Zones: Employing K-Means Clustering in Kolar District, Karnataka. SN Computer Science 5: 1-12.
42. Maione C, Costa N L da, Jr F B, Barbosa R M (2022) A Cluster Analysis Methodology for the Categorization of Soil Samples for Forensic Sciences Based on Elemental Fingerprint. Applied Artificial Intelligence https://www.tandfonline.com/doi/abs/10.1080/08839514.2021.2010941.
43. Boateng E Y, Otoo J, Abaye D A (2020) Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. Journal of Data Analysis and Information Processing 8: 4.
44. Ahmed S F, Alam Md S B, Hassan M, Rozbu M R, Ishtiak T, et al. (2023) Deep learning modelling techniques: Current progress, applications, advantages, and challenges. Artificial Intelligence Review 56: 13521-13617.
45. Odhiambo J M, Mvurya D M, Luvanda D A, Mwakondo D F (n.d.) Deep Learning Algorithm for Identifying Microplastics in Open Sewer Systems: A Systematic Review.
46. Wikle C K, Zammit-Mangion A (2023) Statistical Deep Learning for Spatial and Spatiotemporal Data. Annual Review of Statistics and Its Application 10: 247-270.
47. Sharma S, Beslity J O, Rustad L, Shelby L J, Manos P T, et al. (2024) Remote Sensing and GIS in Natural Resource Management: Comparing Tools and Emphasizing the Importance of In-Situ Data. Remote Sensing 16: 22.
48. Delaine F (2020) In situ calibration of low-cost instrumentation for the measurement of ambient quantities: Evaluation methodology of the algorithms and diagnosis of drifts [Phdthesis, Institut Polytechnique de Paris]. https://theses.hal.science/tel-03086234.
49. Hu T, Lai Q, Fan W, Zhang Y, Liu Z (2023) Advances in Portable Heavy Metal Ion Sensors. Sensors 23: 8.
50. Chuvieco E (2020) Fundamentals of Satellite Remote Sensing: An Environmental Approach, Third Edition (3rd ed.). CRC Press. https://doi.org/10.1201/9780429506482.
51. Lovynska V, Bayat B, Bol R, Moradi S, Rahmati M, et al. (2024) Monitoring Heavy Metals and Metalloids in Soils and Vegetation by Remote Sensing: A Review. Remote Sensing 16: 17.
52. Mehendale N, Neoge S (2020) Review on Lidar Technology (SSRN Scholarly Paper No. 3604309). Social Science Research Network. https://doi.org/10.2139/ssrn.3604309.
53. Esther Darkwah (2023) Developing spatial risk maps of PFAS contamination in farmlands using soil core sampling and GIS. World Journal of Ad

Advanced Research and Reviews 20: 2305-2325.
54. Mohammad Aman Ullah Sunny (2024) Unveiling spatial insights: Navigating the parameters of dynamic Geographic Information Systems (GIS) analysis. International Journal of Science and Research Archive 11: 1976-1985.
55. Sarker I H (2021) Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science 2: 160.
56. Cho B, Dayrit T, Gao Y, Wang Z, Hong T, et al. (2020) Effective Missing Value Imputation Methods for Building Monitoring Data. 2020 IEEE International Conference on Big Data (Big Data) 2866-2875.
57. Jäger S, Allhorn A, Bießmann F (2021) A Benchmark for Data Imputation Methods. Frontiers in Big Data 4.
58. Ali M S, Islam M K, Das A A, Duranta D U S, Haque Mst F, et al. (2023) A Novel Approach for Best Parameters Selection and Feature Engineering to Analyze and Detect Diabetes: Machine Learning Insights. BioMed Research International 2023: 8583210.
59. Katya E (2023) Exploring Feature Engineering Strategies for Improving Predictive Models in Data Science. Research Journal of Computer Systems and Engineering 4: 2.
60. Verdonck T, Baesens B, Óskarsdóttir M, vanden Broucke S (2024) Special issue on feature engineering editorial. Machine Learning 113: 3917-3928.
61. Ishola (2021) Application of satellite based remote sensing to the estimation and monitoring of crop health http://irepo.futminna.edu.ng:8080/jspui/handle/123456789/14492.
62. R Dean J, Ahmed S, Cheung W, Salaudeen I, Reynolds M, et al. (2024) Use of remote sensing to assess vegetative stress as a proxy for soil contamination. Environmental Science: Processes & Impacts 26: 161-176.
63. Borah K, Das H S, Seth S, Mallick K, Rahaman Z, et al. (2024) A review on advancements in feature selection and feature extraction for high-dimensional NGS data analysis. Functional & Integrative Genomics 24: 139.
64. Cheng Y, Wang X, Xia Y (2021). Supervised t-Distributed Stochastic Neighbor Embedding for Data Visualization and Classification. INFORMS Journal on Computing 33: 566-585.
65. Jia X, Hu B, Marchant B P, Zhou L, Shi Z, et al. (2019) A methodological framework for identifying potential sources of soil heavy metal pollution based on machine learning: A case study in the Yangtze Delta, China. Environmental Pollution 250: 601-609.
66. Lu X, Du J, Zheng L, Wang G, Li X, et al. (2023) Feature fusion improves performance and interpretability of machine learning models in identifying soil pollution of potentially contaminated sites. Ecotoxicology and Environmental Safety 259: 115052.
67. Movahedi F, Padman R, Antaki J F (2023) Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores. The Journal of Thoracic and Cardiovascular Surgery 165: 1433-1442.e2.
68. Singh K P, Gupta S, Rai P (2013) Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment 80: 426-437.
69. Hodson T O (n.d.). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not.
70. Willmott C J, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 30: 79-82.
71. Wang Q, Li C, Hao D, Xu Y, Shi X, et al. (2023) A novel four-dimensional prediction model of soil heavy metal pollution: Geographical explanations beyond artificial intelligence “black box.” Journal of Hazardous Materials 458: 131900.
72. Mallick J, Alqadhi S, Hang H T, Alsubih M (2024) Interpreting optimised data-driven solution with explainable artificial intelligence (XAI) for water quality assessment for better decision- making in pollution management. Environmental Science and Pollution Research 31: 4294- 42969.
73. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, et al. (2023) Explainable AI (XAI): Core Ideas, Techniques, and Solutions. ACM Comput Surv 194: 1-194.
74. Parisineni S R A, Pal M (2024) Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations. International Journal of Data Science and Analytics 18: 457-466.
75. Wani A K, Rahayu F, Ben Amor I, Quadir M,

Murianingrum M, et al. (2024) Environmental resilience through artificial intelligence: Innovations in monitoring and management. Environmental Science and Pollution Research 31: 18379-18395.

76. Sheykhmousa M, Mahdianpari M, Ghanbari H, Mohammadimanesh F, Ghamisi P, et al. (2020) Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13: 6308-6325.
77. Yao L, Xu M, Liu Y, Niu R, Wu X, et al. (2024) Estimating of heavy metal concentration in agricultural soils from hyperspectral satellite sensor imagery: Considering the sources and migration pathways of pollutants. Ecological Indicators 158: 111416.
78. Gong Y, Liu G, Xue Y, Li R, Meng L (2023) A survey on dataset quality in machine learning. Information and Software Technology 162: 107268.
79. Liao T, Taori R, Raji I D, Schmidt L (2021) Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). https://openreview.net/forum?id=mPducS1MsEK.
80. Thiyagalingam J, Shankar M, Fox G, Hey T (2022) Scientific machine learning benchmarks. Nature Reviews Physics 4: 413-420.
81. Ogwu M C (2025) Science and Theory of Pollution: Sources, Pathways, Effects and Pollution Credit. In M. C. Ogwu & S. Chibueze Izah (Eds.), Evaluating Environmental Processes and Technologies Springer Nature Switzerland 117-147.
82. Siddique I (2024) Machine learning empowered geographic information systems: Advancing Spatial analysis and decision making (SSRN Scholarly Paper No. 4892563). Social Science Research Network. https://papers.ssrn.com/abstract=4892563.
83. Choi Y (2023) GeoAI: Integration of Artificial Intelligence, Machine Learning, and Deep Learning with GIS. Applied Sciences 13: 6.
84. Ambasht A (2023) Real-Time Data Integration and Analytics: Empowering Data-Driven Decision Making. International Journal of Computer Trends and Technology 71: 8-14.
85. Mahdavifar S, Ghorbani A A (2019) Application of deep learning to cybersecurity: A survey. Neurocomputing 347: 149-176.