Review

A Systematic Review of Machine Learning Algorithms for Soil Pollutant Detection Using Satellite Imagery

Amir TavallaieNejad * , Maria Cristina da Costa Vila , Gustavo Paneiro , João Santos Baptista

Posted Date: 3 March 2025

doi: 10.20944/preprints202503.0068.v1

Keywords: soil pollutant; satellite imagery; Artificial Intelligence; Machine Learning; pollution detection

 doi:10.20944/preprints202503.0068.v1

Systematic Review

A Systematic Review of Machine Learning Algorithms for Soil Pollutant Detection Using Satellite Imagery

Amir TavallaieNejad [1,*], Maria Cristina da Costa Vila [1], Gustavo Paneiro [2] and João Santos Baptista [1]

[1] CERENA, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

[2] CERENA/DER, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal

* Correspondence: up202111331@up.pt; Tel.: +13688820503

Abstract: Soil preservation from pollutants is essential for sustaining human and ecological health. This review explores the application of satellite imagery and machine learning (ML) techniques in detecting soil pollution, addressing recent advancements and key challenges in this field. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, a comprehensive search across three major databases yielded 47 articles from an initial pool of 1,018 publications spanning the last eight years. Among these, 34 studies focused on direct detection of soil pollutants, while 13 examined relationships between vegetation indicators and soil contaminants. This review evaluates various satellite platforms, highlights limitations of existing spaceborne sensors, and compares the effectiveness of ML models for soil pollution detection. Key challenges include the lack of standardization in datasets and methodologies, variations in evaluation metrics, and differences in algorithmic performance across studies. Findings emphasize the need for standardized frameworks and improved sensor capabilities to enhance detection accuracy. This work provides a foundation for future research, encouraging the integration of advanced ML models and multi-sensor satellite data for comprehensive soil pollution monitoring.

Keywords: soil pollutant; satellite imagery; Artificial Intelligence; Machine Learning; pollution detection

1. Introduction

Soil, a non-renewable resource and fundamental component of ecosystems, plays a vital role in supporting diverse ecological processes and sustaining life on Earth [1]. However, soil integrity is increasingly threatened by contaminants, including heavy metals, organic compounds from oil spills, and microplastics, which degrade soil quality and pose significant environmental and human health risks [2]. Contaminated soil can enter the human body through ingestion, inhalation, dermal contact, and indirect pathways, raising concerns about potential adverse health effects [3]. Soil pollution arises from both synthetic chemical introductions and natural alterations to soil properties [3]. Common sources of contamination include ruptured subsurface storage tanks, pesticide applications, surface water percolation, fuel and oil disposal, landfill leachate, and direct industrial discharge [3]. Among concerning pollutants are potentially harmful elements (PHEs), including heavy metals, semi-metals, and non-metals. These elements, like copper (Cu), lead (Pb), mercury (Hg), cadmium (Cd), and zinc (Zn), are naturally present at low levels due to weathering of parent materials [4]. Under acidic soil conditions, PHEs can become soluble, impacting plants and contaminating groundwater [5]. Pollution by PHEs is particularly concerning in agricultural areas, where it threatens food security and environmental sustainability [6,7]. Alarmingly, approximately five million cases of soil

 doi:10.20944/preprints202503.0068.v1

contamination by metals or metalloids have been documented globally, affecting an estimated 506 million hectares of land [8].

Traditional soil sampling methods, which rely on discrete sampling, provide limited data points for constructing continuous maps of soil contamination, making it difficult to assess pollution levels over large geographic areas [9]. Laboratory analyses of soil samples, especially on a large scale, are time-intensive, costly, and environmentally burdensome [10]. Remote sensing technology, particularly visible and near-infrared reflectance (VNIR) spectroscopy, has thus emerged as a promising, scalable solution for cost-efficient monitoring of soil contamination [11].

In recent years, research has increasingly focused on connections between hyperspectral data and soil contaminants. Numerous machine learning approaches, primarily regression-based, have been applied at field scales using laboratory-processed soil data and hyperspectral measurements [9,12–19]. These methods have shown effectiveness in detecting various heavy metals, including chromium (Cr), arsenic (As), lead (Pb), cadmium (Cd), and zinc (Zn). For instance, Zhang et al. [16] analyzed spectral reflectance in relation to concentrations of Cr, As, nickel (Ni), and Cd, achieving strong correlation coefficients for As and Cd (0.456 and 0.548) using Continuous Wavelet Transform (CWT). Liu et al. [18] achieved an 81.1% correlation by integrating backpropagation neural networks (BPNN) with particle swarm optimization (PSO), forming the Particle Swarm Optimized-Back Propagation Neural Network (PSO-BPNN) for detecting arsenic.

The availability of satellite data and advancements in Graphics Processing Unit (GPU) capabilities [20] have enabled the application of deep learning methods, which outperform traditional machine learning models in various remote sensing tasks, including estimating precipitation, soil moisture, land surface and air temperature, crop yield, and pollutant detection [21–25]. Recurrent neural networks (RNNs) such as long short-term memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM), and gated recurrent units (GRU) have been developed to capture temporal relationships in time-series data, showing promise for satellite-based evaluations [22,23].

Despite the expanding literature on soil pollution detection using satellite imagery, the field continues to evolve rapidly, driven by increasingly accessible and accurate satellite data. Several satellites, notably Sentinel-2 and Landsat-8, are frequently used for soil contamination studies [24]. A variety of machine learning methods have been applied, with Random Forest (RF) being the most popular, followed by techniques such as Cubist, Partial Least Squares Regression (PLSR), and Support Vector Machine (SVM) [25].

Evaluating detection results is critical in this field, with each study employing one or more metrics to assess model performance. Commonly used metrics include R-squared (R^2), Root Mean Square Error (RMSE), overall accuracy, F1 score, kappa statistics, and Ratio of Prediction to Deviation (RPD) [26].

In this evolving landscape of soil pollution detection, where advancements in machine learning (ML) and remote sensing drive progress, a review that consolidates existing evidence and insights is timely. While systematic reviews are common in life sciences, they are less frequently applied in earth and environmental sciences. This review synthesizes the literature on machine learning techniques for detecting soil pollutants through satellite imagery, offering new perspectives and insights into methodologies, data sources, and evaluation metrics. By conducting a rigorous analysis, this review not only bridges knowledge gaps but also advances the field of remote sensing technology for soil pollution monitoring. This work contributes not just a compilation of insights but a fresh vantage point that accelerates the progression of this critical field.

2. Materials and Methods

The methodology followed the PRISMA Extension Guidelines (PRISMA ScR; [27–29]). A checklist is available in Figure 1 within Supplementary Material Appendix A. This section explains the Eligibility Criteria, Characteristics of Accepted Studies (Participants, Detections Considered and Evaluated, and Study Design), Information Sources, Search Strategy, and additional data management and selection processes.

 doi:10.20944/preprints202503.0068.v1

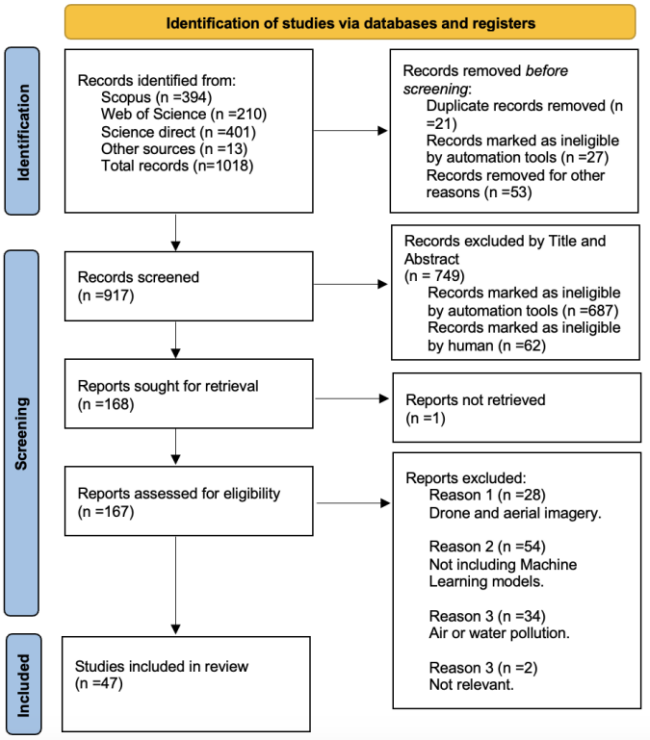

Figure 1. PRISMA Flow diagram of the research [27].

2.1. Eligibility Criteria: Search Characteristics

This systematic three-stage search approach ensured a comprehensive and focused collection of pertinent literature, enhancing the reliability and relevance of the research findings.

The review of soil pollution detection methods using satellite imagery and machine learning followed a structured, three-phase process. In the initial phase, a comprehensive search of literature from 2016 to 2024 was conducted, using predefined keywords based on pertinent subject definitions. The search was limited to articles, cohort studies, case reports, and cross-sectional studies published in indexed journals in English, drawn from 17 peer-reviewed journals.

In the second phase, articles were meticulously selected based on eligibility criteria to ensure alignment with the research objectives. Only those meeting all criteria were advanced for further analysis. In the final phase, abstracts of the selected articles were reviewed to assess relevance to the research topic, including only those closely aligned with the study focus. This review approach ensured a comprehensive, relevant collection of literature.

2.2. Characteristics of Accepted Studies

Participants

The investigation for relevant research in the field of Machine Learning applied to soil contaminants was performed without imposing any specific restrictions on the type of contaminants or soil types. The search encompassed studies that employed Machine Learning algorithms on various types of satellite images.

The inclusion criteria were deliberately broad to ensure a comprehensive representation of research in this domain. Consequently, all studies involving Machine Learning techniques on soil contaminants, regardless of the specific contaminants involved or the type of soil, were considered.

Furthermore, studies utilizing diverse satellite images and employing various Machine Learning algorithms were included in the search. This approach aimed to encompass the broad spectrum of methodologies employed in this field, providing a comprehensive and diverse collection of studies for analysis.

 doi:10.20944/preprints202503.0068.v1

By adopting an inclusive approach to participant selection and study characteristics, this research ensures a comprehensive and well-rounded exploration of Machine Learning applications on soil contaminants using satellite imagery.

Detections Considered and Evaluated

The study focused on the following detection categories:

- **Assessment of satellite imagery:** Analysis of images captured by satellites.
- **Machine learning methodologies:** Evaluation of machine learning techniques used for satellite image analysis.
- **Identification of contaminated soil types**: Detection and assessment of contamination types through analysis described in (B), evaluated by (A).
- **Validation of results:** Verification of identified contaminated soil types (C) by comparing with on-ground sampling data.

By evaluating these categories, the study aimed to provide a comprehensive understanding of machine learning applications for soil contaminants.

Design of Accepted Studies

The design of accepted studies focused on specific criteria, including the use of satellite images, computational analysis, and image processing to assess soil characteristics. The objective was to develop models with minimal error and high accuracy, enabling generalization to various locations and contaminations. Notably, contaminants included heavy metals, chemicals, and soil-related features like moisture, erosion, salinity, and evaporation. Studies using images from aircraft or drones were excluded to emphasize the unique advantages of satellite imagery in soil characterization and pollution detection.

2.3. Information Sources

Data was collected from key digital databases, including ScienceDirect, Scopus, and Web of Science. Filters applied within each database included publication year (2016–2024), document type (articles and articles in press), and language (English). Only scholarly journal articles and scientific publications were considered.

2.4. Search Strategy

The search strategy comprises the amalgamation of the subsequent keywords: “Machine Learning”, “Deep learning”, “Artificial Intelligence”, “Polluta*”, “Contamina*”, Satellite, Algorithm, Image, and Soil.

Consequently, thirteen combinations were deemed valid:

1. “Machine learning” AND polluta* AND Satellite;
2. “Deep learning” AND polluta* AND Satellite;
3. “Artificial Intelligence” AND polluta* AND Satellite;
4. “Machine learning” AND polluta* AND image AND soil;
5. “Deep learning” AND polluta* AND image AND soil;
6. “Artificial Intelligence” AND polluta* AND image AND soil;
7. “Machine learning” AND polluta* AND Algorithm AND Satellite;
8. “Machine learning” AND Soil contamina* AND Satellite;
9. “Deep learning” AND contamina* AND Satellite AND soil;
10. “Artificial Intelligence” AND Soil contamina* AND Satellite;
11. “Machine learning” AND Soil AND contamina* AND image;
12. “Deep learning” AND contamina* AND image AND soil;
13. “Artificial Intelligence” AND Soil AND contamina* AND image.

 doi:10.20944/preprints202503.0068.v1

2.5. Study Records

2.5.1. Data Management

For this study, data management involved obtaining information directly from academic databases. Studies were systematically archived using Mendeley Reference Manager to ensure orderly arrangement and ease of retrieval.

Following identification, data extraction was conducted using a customized Excel table. Each row represented an individual record, while each column captured specific parameters from each article. This structured approach facilitated accurate data compilation from the sources.

By using reference management software and a structured Excel table, data was efficiently organized and made ready for detailed analysis of research findings.

2.5.2. Selection Process

The selection process involved multiple phases to ensure the inclusion of relevant and eligible studies for the review. Initially, records were automatically screened based on predefined criteria, including publication year (2016–2024, to align with the PRISMA method's focus on recent literature), document type (articles and forthcoming articles), source type (peer-reviewed journals), and language (English). This automated screening helped narrow the pool of potential studies.

Next, a researcher conducted a manual review to assess each work's alignment with the research objectives, based solely on screening titles and abstracts. Any records with uncertain relevance after this initial verification proceeded to the next selection phase.

In the second screening stage, the methodologies used in each study were evaluated against the eligibility criteria, requiring a comprehensive reading of each article. Studies accepted for inclusion met all of the following minimum criteria:

- Utilization of Artificial Intelligence techniques, such as Machine Learning (ML) and Deep Learning (DL), for image processing.
- Detection of soil pollutants or soil characteristics related to soil pollutants, such as soil evaporation or soil moisture content.
- Exclusive use of satellite images (excluding airborne or drone data and soil testing data) for analysis.
- Validation of data through methods that measure accuracy or error in detecting soil pollutants, such as field soil testing.
- Description and specification of the Machine Learning methods used for image processing, especially if developed by the author.
- Defined accuracy metrics and presentation of results in quantifiable terms.

By adhering to these stringent selection criteria, this study ensured the inclusion of works that met the research objectives and contributed to a comprehensive analysis of Artificial Intelligence applications in soil pollutant detection using satellite imagery.

2.6. Data Collection Process

During data collection, the results of each study were presented in a detailed description using a table format. Both the results and conclusions were carefully analyzed, focusing on content with potential causal links to the objectives of the systematic review.

The information extracted and included in Appendix A comprised the following key elements:

- **References:** Each article reference is provided. Following a meticulous review of over 1,000 articles using the PRISMA method, a final selection of 36 articles was made to align with the specific goals of this review. Further details on these selected articles are presented in the accompanying table.

 doi:10.20944/preprints202503.0068.v1

- **Satellite Name:** The appendix's second column lists the satellites used in each article, with some studies focusing on imagery from a single satellite while most compare multiple satellites. This approach aids in selecting the most suitable satellite(s) for specific applications.
- **Machine Learning Methods:** The types of Machine Learning (ML) techniques employed by the authors and any unique methodologies they developed.
- Contaminant Types: A list of all contaminants or pollutants studied, as well as other soil characteristics related to soil pollutants. The names of these parameters are presented in the fourth column of the appendix.
- **Validation:** Validation of each ML method is a critical aspect, evaluating the effectiveness of the techniques for interpreting satellite images. Typically, this validation involves direct soil sampling at specified depths, ensuring an understanding of the correlation between soil samples and satellite imagery. The number of boreholes used varies with the investigated area, as indicated in the appendix.
- **Performance:** Machine learning performance metrics serve as quantitative measures to assess model accuracy and effectiveness across classification, regression, or clustering tasks. Each article's specific performance metrics and the best results are summarized in a table or presented as final values in the appendix.
- **Results:** The last column of Appendix A provides a concise conclusion for each article, showing how each study's results align with the systematic review's purpose. This section facilitates an understanding of the effectiveness of different methods, satellite types, and other parameters for pollution detection.

By systematically gathering and organizing this comprehensive data, the review enabled an in-depth analysis and drew meaningful conclusions on the application of Machine Learning for soil pollutant detection using satellite imagery.

2.7. Prioritization and Outcomes

Quantitative Prioritization:

- Validation Sample Size: Studies using larger sample sizes for validation were given higher priority. Substantial sample sizes enhance the statistical robustness and reliability of findings, improving the overall quality of the research.
- Accuracy of Validation Data: Articles that employed the most accurate and reliable validation methods for analyzing images through Machine Learning were prioritized. High-quality validation data adds credibility to research outcomes and strengthens confidence in the reported results.

Qualitative Prioritization:

- Identification of Different Contaminations: Studies that directly identified various contaminations or indirectly inferred them through specific soil characteristics were considered more significant. This approach broadens the research scope, providing insights into a diverse range of soil pollutants and their potential impacts.
- Diverse ML Methods: Articles exploring various Machine Learning methods for image analysis were prioritized. Employing diverse ML methodologies enables a comprehensive exploration of soil pollutant detection approaches, enhancing understanding of their effectiveness.

By employing this comprehensive prioritization approach, the review aimed to highlight studies with strong methodological rigor, broad scope, and substantial potential to provide valuable insights into the use of Machine Learning for soil pollutant detection using satellite imagery.

2.8. The Risk of Bias and Quality Assessment

The risk of bias was assessed at the study level, considering various methodological and data analysis parameters (see Table 1). These parameters included satellite source diversity, variations in ML methodologies, contaminant diversity, performance analysis methods, sampling quantity and

 doi:10.20944/preprints202503.0068.v1

quality, and validation data. Each parameter's potential impact on study outcomes was classified as "high risk," "low risk," or "unclear risk" according to established criteria [30].

Table 1. Bias analysis of each study.

Study	Different Satellite sources	ML Methodology Variation	Variation of Detected Contaminant	performance Analysis Alteration	Sampling Quantity	Sampling Quality	Validation data
[31]	LR	HR	LR	LR	LR	LR	LR
[32]	LR	LR	UR	HR	LR	LR	LR
[33]	HR	HR	LR	LR	UR	UR	LR
[34]	HR	LR	HR	LR	UR	LR	LR
[35]	LR	HR	UR	LR	LR	LR	LR
[36]	HR	HR	UR	LR	UR	HR	LR
[37]	HR	LR	LR	LR	LR	LR	LR
[38]	LR	LR	UR	LR	HR	LR	LR
[39]	HR	LR	UR	UR	LR	LR	LR
[40]	LR	LR	HR	LR	LR	LR	LR
[41]	HR	HR	UR	LR	HR	LR	LR
[42]	HR	LR	HR	LR	LR	LR	LR
[43]	HR	HR	UR	UR	HR	LR	UR
[44]	HR	LR	UR	LR	LR	LR	LR
[45]	HR	LR	HR	LR	LR	LR	LR
[46]	HR	LR	LR	LR	LR	LR	LR
[47]	HR	LR	LR	LR	LR	LR	LR
[48]	LR	LR	LR	LR	LR	LR	LR
[49]	HR	HR	UR	UR	UR	UR	LR
[50]	HR	HR	UR	HR	HR	LR	LR
[51]	HR	LR	LR	LR	LR	LR	LR
[52]	LR	HR	LR	HR	HR	LR	LR
[53]	HR	LR	UR	LR	LR	LR	LR
[54]	LR	LR	LR	LR	LR	LR	LR
[55]	LR	LR	LR	LR	LR	LR	LR
[56]	HR	LR	LR	HR	LR	LR	LR
[57]	LR	HR	UR	LR	LR	LR	LR
[58]	HR	LR	HR	LR	HR	LR	LR
[59]	HR	LR	UR	LR	LR	LR	LR
[60]	LR	LR	UR	LR	UR	LR	LR
[61]	LR	HR	UR	LR	UR	LR	LR
[62]	LR	HR	UR	HR	LR	LR	LR
[63]	LR	HR	UR	LR	HR	LR	LR
[64]	LR	HR	UR	LR	LR	LR	LR
[14]	LR	LR	HR	LR	LR	LR	LR
[65]	HR	LR	LR	LR	LR	LR	LR

HR: high risk; LR: low risk; UR: unclear risk.

The risk of bias analysis in each study examined the following specific items:

- **Different Satellite Sources:** Evaluation of the satellite sources used for image data in each study.
- **ML Methodology Variation:** Assessment of the variety in machine learning methods used for image analysis.
- **Variation of Detected Contaminants:** Analysis of the different types of soil contaminants detected through satellite imagery.
- **Performance Analysis Alteration:** Evaluation of the performance metrics used to assess ML methods' efficiency in pollution detection.
- **Sampling Quantity:** Consideration of sample size, as data quantity influences ML performance and evaluation reliability.

 doi:10.20944/preprints202503.0068.v1

- **Sampling Quality:** Assessment of sample quality, including adherence to standards and accurate classification of pollutants or soil characteristics.
- **Validation Data:** Evaluation of the accuracy and effectiveness of methods used to detect soil characteristics from satellite images.

With the exception of sampling quality and quantity, all other variables in the methodology section emphasize the need for improved reporting regarding variation and validation of results.

2.9. Article Selection

Following PRISMA guidelines [66], the initial database searches yielded 1,018 items. After removing duplicates and ineligible entries using automated tools and other criteria, 917 records remained for further consideration. During the screening process, 749 records were excluded based on title and abstract ineligibility, identified through a combination of automated and manual assessment. Additional exclusions were made for studies involving drone and aerial imagery, lacking machine learning models, or unrelated to soil pollution, resulting in a final selection of 47 articles for detailed review. Figure 1 provides an overview of the number of articles identified at each step, following the PRISMA methodology.

3. Results, Discussion, and Summary of Evidence

This section presents the culmination of a systematic review, exploring the advancements in satellite-based soil contaminant detection. The investigation leverages satellite imagery and includes the selection and analysis of relevant studies employing state-of-the-art Machine Learning methodologies. Through meticulous screening and evaluation, the efficacy of satellite systems in detecting diverse soil pollutants and characteristics is revealed. Emphasizing cost-effectiveness, accessibility, and accuracy, this review highlights the transformative potential of satellite-based methods in soil pollution monitoring and management.

Satellite imagery has emerged as an effective method for soil contaminant detection, offering advantages in cost, accessibility, and accuracy. In the pursuit of identifying soil contaminants, researchers have utilized data from several satellites, often incorporating images from multiple sources in their studies. Among the most widely used satellites, Sentinel-2 [14,31,35,39,40,47,50,52,56,60,61,63–65,67–69] and Landsat 8 [14,31,33,34,36,40,47,50,53,54,60,62,63,64,70–73] stand out, with each appearing frequently in the literature. These satellites, integral to distinct Earth observation systems, have been applied across various domains, including land-use mapping, environmental monitoring, and natural resource management. Specifically, seven studies used Sentinel-2 for soil property detection [40,56,60,61,63,64,74] and four studies employed it for heavy metal detection [14,31,47,52]. Additionally, five studies utilized Sentinel-2 for vegetation property assessment [39,50,67–69].

For Landsat 8, seven studies focused on heavy metal detection [14,34,53,54,71,72,75], another seven on soil property detection [40,60,62,63,70,73,76], and one study assessed vegetation properties [36]. Sentinel-2, part of the Copernicus Programme developed by the European Space Agency (ESA), provides high-resolution, multispectral images of the Earth's surface. Equipped with a multi-spectral imaging instrument that captures data in 13 spectral bands—from visible to shortwave infrared—the satellite achieves a spatial resolution between 10 to 60 meters and covers the Earth's land surface every five days [58,77].

Landsat 8, developed through the collaborative efforts of NASA and the United States Geological Survey (USGS) under the Landsat program, is equipped with two instruments: the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS). OLI captures data in nine spectral bands, covering the visible to shortwave infrared spectrum, with a spatial resolution ranging from 15 to 100 meters. TIRS captures data in two thermal infrared bands with a 100-meter resolution, providing comprehensive global land coverage every 16 days [78].

 doi:10.20944/preprints202503.0068.v1

Nearly 40 percent of the reviewed articles (18 out of 47) utilized multiple satellites, allowing comparisons across satellite systems and creating a robust dataset for analysis [14,31,32,35,40,44,48,52,54,56,60,61,62–64,67,68,72]. This multi-satellite approach enhances data reliability and enables a more comprehensive analysis of soil contaminants. By combining data from different satellite sources, researchers can address limitations inherent to individual satellite systems, such as temporal resolution and spectral range, thereby improving overall accuracy and analytical depth.

In summary, the use of satellite imagery for soil contaminant detection is increasingly validated by a growing body of literature. The integration of multiple satellite datasets and advanced Machine Learning techniques holds significant potential for improving soil pollution monitoring and management. As technology advances and more sophisticated satellites are launched, the precision and applicability of these methods are expected to enhance further, paving the way for more effective environmental conservation strategies.

The analysis of machine learning (ML) methods for soil contaminant detection reveals distinct patterns and preferences. The most frequently used methods are decision-tree-based models, as illustrated in Figure 2, including Random Forest (RF), ExtraTrees (ET), Decision Tree, Cubist (Cu), Classification and Regression Trees (CART), and the Deep Forest Algorithm. Among these, Random Forest emerges as the most prevalent, with 33 instances in the literature [20,37,40–59,67–73,75,76,79,80]. This preference underscores the robustness and effectiveness of Random Forest in handling the complexities of soil contaminant detection.

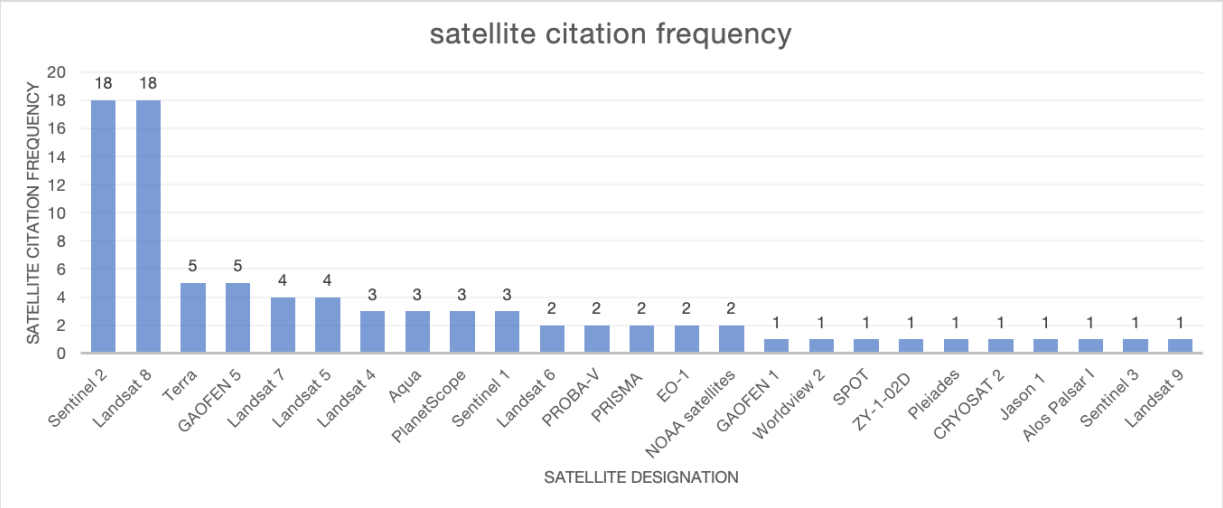

Figure 2. Satellite citation frequency studied in the selected works.

Random Forest is well-regarded for its proficiency in land cover classification and feature extraction. Studies have applied RF to classify land cover types from Sentinel-2 imagery, achieving an overall accuracy of 83% [50]. Similarly, Chen et al. used RF to extract urban green spaces from Landsat 8 imagery, obtaining an accuracy of 89.15% [44]. RF's strengths in satellite image processing include its ability to handle large datasets with high-dimensional features and robustness in managing noise and outliers. Additionally, the model allows for feature importance analysis, which aids in identifying crucial variables for classification and feature extraction tasks [45].

Following Random Forest, other frequently used methods include Support Vector Machine (SVM) with 10 instances [38,39,46,48,55,58,65,68,76,80], Cubist with six instances [37,54,57,59,64,75], and both Artificial Neural Network (ANN) and Partial Least Squares Regression (PLSR) with six instances each [14,33,34,40,58,79].

SVM, a versatile supervised learning algorithm, is commonly used for classification and regression. It works by finding an optimal hyperplane that separates data classes. SVM has been successfully applied in remote sensing and image analysis, achieving reliable classification results

 doi:10.20944/preprints202503.0068.v1

based on satellite imagery [37,75]. Its versatility and ability to handle complex datasets make SVM valuable in soil contaminant detection [57].

Other methods include KNN [37,46,56,58,68], BPNN [47,55,58], CART [39,46,56], CNN [43,48,52,67,70], MLP [46,58,67,73], ADB [42,46], ELM [40,55,70], LSTM [32,52,80], Pix RNN [36,52], and SVR [34,37,68]. The varied frequency of these methods reflects a diverse approach to satellite imagery analysis for soil contaminant detection, with each method contributing unique strengths and capabilities.

dditionally, 56% (32 out of 57) of the reviewed articles used multiple ML models to compare and validate datasets, which allowed researchers to identify the most effective model for specific environmental parameters [14,32,34,36–40,44–48,52–60,67,68,70,71,73,75,76,79–81]. Results showed that Random Forest was most often (13 of 32 instances) identified as the best model [38,45,47,48,53–56,59,68,71,75,81]. while SVR and ANN each achieved the best results three times.

The reliance on Random Forest illustrates its robustness and reliability in soil contaminant detection. The diversity of other methods highlights the tailored approaches researchers adopt to address specific challenges. As ML techniques advance and integrate with new satellite technologies, the accuracy and applicability of these methods are expected to improve, enhancing soil contamination monitoring and environmental management.

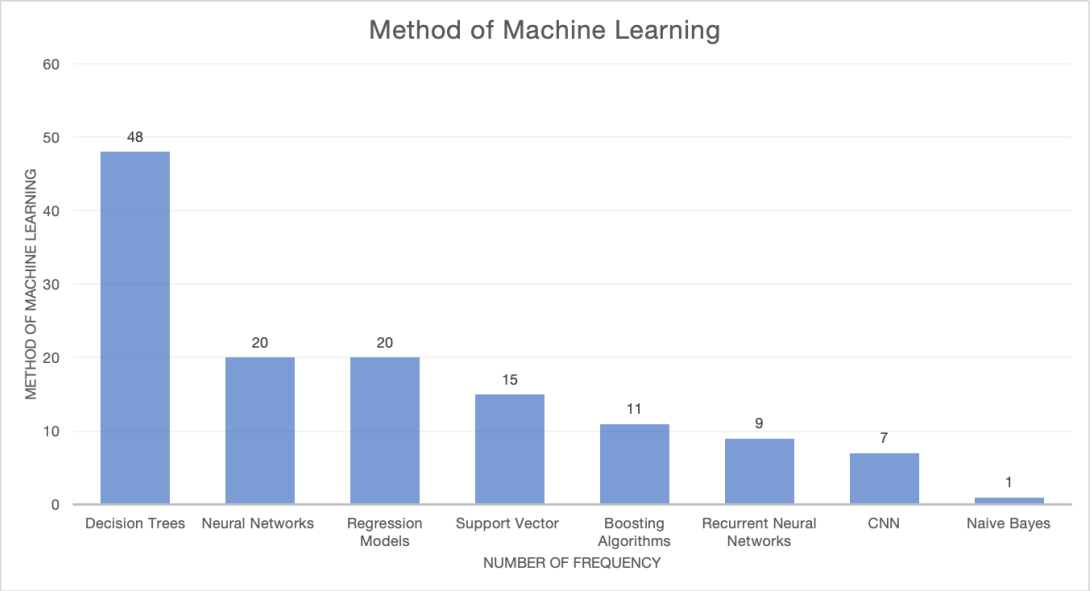

Figure 3. Types of Machine Learning models studied in the selected works.

Evaluating detection results is crucial in this research, with each article employing one or more evaluation metrics. Among the selected studies, 30 used Root Mean Squared Error (RMSE) [14,32,34,37,38,40,42,44,45,46,54,55,57–61,63–65,68,69,71–73,75,79,80]. a metric that measures the average difference between predicted and actual values. For instance, Ma et al. [32] used RMSE to evaluate a deep learning model for image super-resolution, achieving an RMSE of 9% on the Actin validation set [70].

The equation of Root Mean Squared Error (RMSE) is as follows:

$$\boldsymbol{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - y_p)^2}{n}}, \quad (1)$$

where y_i represents the actual (observed) value of the target variable for data point i, and y_p represents the predicted value of the target variable for data point i.

In addition to RMSE, 25 articles used the coefficient of determination (R^2) [14,31,32,34,37,38,40,44–47,55,57–60,63,64,68–73,75,79]. indicating how well the model fits the data. The equation of the coefficient of determination (R^2) is given by:

 doi:10.20944/preprints202503.0068.v1

$$R^2 = 1 - \frac{\sum(y_i - y_p)^2}{\sum(y_i - \bar{y})^2}, \quad (2)$$

Where $\bar{y}$ is the mean of the actual values.

Moreover, 16 articles reported overall accuracy [35,36,38,39,45,48,50,52,53,56,58,62,67,70], with other metrics such as Mean Absolute Error (MAE)[31,34,37,42,55] and Ratio of Prediction to Deviation (RPD) [14,46,47,64,65] (Figure 4).

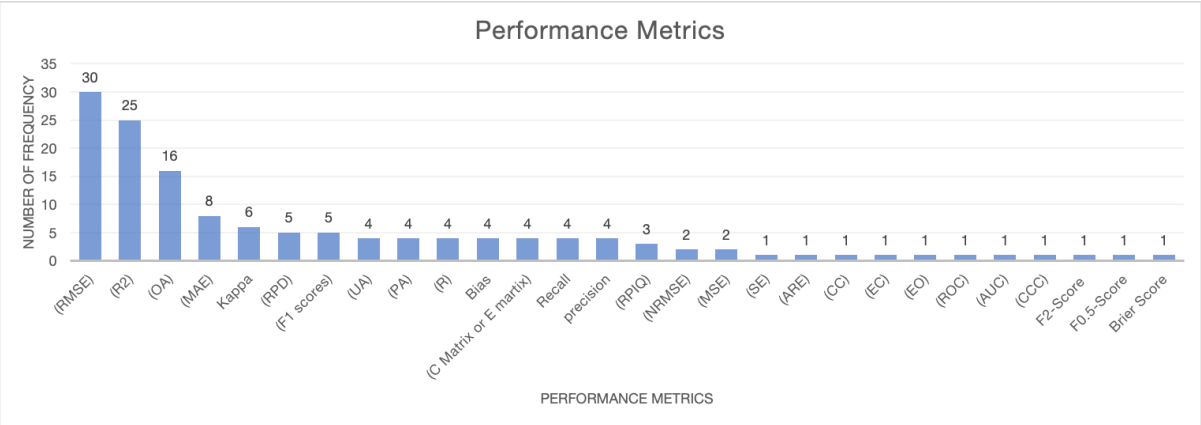

Figure 4. Performance metrics studied in the selected works.

The primary objective of this study was to identify environmental parameters associated with soil pollution, both directly and indirectly. Notably, vegetation properties were linked to specific types of pollution, impacting various vegetation types [32]. Among the parameters examined, heavy metals played a crucial role, focusing on copper (Cu), arsenic (As), cadmium (Cd), chromium (Cr), nickel (Ni), lead (Pb), zinc (Zn), and iron (Fe). These metals were assessed based on three core factors: gravity, toxicity, and mobility. The gravity factor referred to the extent and concentration of heavy metal contamination in the soil, while the toxicity factor measured harmful effects on human health, plants, and animals. The mobility factor gauged each metal's potential to migrate from soil to other environmental compartments, such as groundwater or surface water, potentially causing contamination in these areas [82].

Figure 5 shows that heavy metals were the most frequently studied environmental parameters, with a frequency count of 47—substantially higher than other pollutants. Specifically, copper (Cu) was examined eight times [34,46,53–55,71,75,79], chromium (Cr) six times [31,37,46,53,54,71], iron (Fe) six times [31,37,47,53,75], arsenic (As) seven times [42,46,53,54,58,71,79] nickel (Ni) appeared five times [53–55,71,75], zinc (Zn) six times [31,37,45,53,55,71,79], lead (Pb) six times [37,47,54,71,72,79], and cadmium (Cd) two times [47,53]. Following heavy metals, soil characteristics were the second most frequently studied environmental parameters, appearing 15 times. These characteristics included soil organic carbon, soil organic matter (SOM), surface soil moisture, soil loss, soil erosion, soil texture, clay content, soil pH, soil salinity, and evaporation [38,40,41,44,47,49,53,56,57,59,61–65].

 doi:10.20944/preprints202503.0068.v1

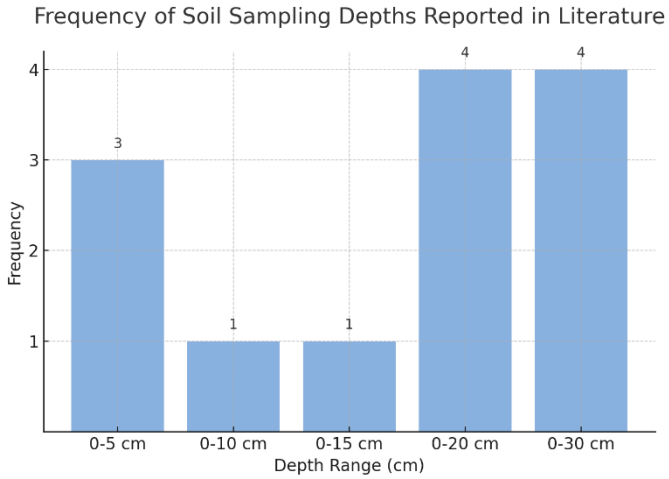

Figure 5. Number of Articles Mentioning Soil Sampling Depth.

Plastic pollution, including microplastics and various plastic materials such as polyethylene (PE), polypropylene (PP), polyvinyl chloride (PVC), polyethylene terephthalate (PET), polystyrene (PS), acrylonitrile butadiene styrene (ABS), ethylene vinyl acetate (EVA), polyamide (PA), polycarbonate (PC), and polymethyl methacrylate (PMMA), ranked third in frequency in the selected articles, appearing 11 times [43,48]. These findings underscore the significance of heavy metals and soil characteristics in soil pollution studies, as well as the rising concern over plastic pollution and its environmental impact.

Satellite images have been instrumental in detecting direct and indirect indicators of soil pollutants across studies. As shown in Table 2, heavy metals were directly identified in 34 articles through satellite imagery, while 15 articles focused on detecting a range of soil parameters, including soil organic carbon, SOM, surface soil moisture, soil loss, soil erosion, soil texture, clay content, soil pH, soil salinity, and evaporation.

Additionally, plastic pollution, which encompasses microplastics and specific plastic types such as PE, PP, PVC, PET, PS, ABS, EVA, PA, PC, and PMMA, is also detectable through satellite imagery. Vegetation properties, including land cover, leaf area index (LAI), cropland suitability, and crop productivity, have emerged as indicators of soil pollutants when analyzed using satellite data. This comprehensive overview highlights the diverse range of environmental parameters that satellite technology can effectively monitor for soil pollution assessment and identification.

Table 2. Frequency of Machine Learning Methods Used in References.

Method of Machine Learning	Number of frequency
Decision Trees (Random Forest (RF), ExtraTrees (ET), Decision Tree, Cubist (Cu), Classification and Regression Trees (CART), Deep Forest Algorithm)	48
Neural Networks (Artificial Neural Network (ANN), Backward Propagation Neural Network (BPNN), The general regression neural network (GRNN), Patch-based multi-image NN system (Patch multi), Patch-based single-image NN system (Patch Single), Pixel-based multi-image NN system (Pix multi), Pixel-based single-image NN system (Pix single), Multi-Layer Perceptron (MLP), Extreme Learning Machine (ELM))	20
Regression Models (Partial Least Square Regression (PLSR), K Nearest Neighbor (KNN), Gaussian Process Regression (GPR), Generalised Linear Models (GLM), Geographically Weighted Regression (GWR), kernel ridge regression (KRR), Linear	20

 doi:10.20944/preprints202503.0068.v1

Regression Model, Stepwise Multiple Linear Regression (SMLR), Multiple Linear Regression (MLR))	
Support Vector (Support Vector Machine (SVM), support vector regression (SVR))	15
Boosting Algorithms (Gradient Descent Boosting Trees (GDB), Adaptive Boosting (ADB), eXtreme Gradient Boosting (XGB), Generalised Boosting Methods (GBM))	11
Recurrent Neural Networks (RNNs) (Long short-term memory (LSTM), Bidirectional LSTM (Bi-LSTM), Gated Recurrent Unit (GRU), Pixel-based RNN system (Pix RNN), Proposed patch-based RNN (PB-RNN))	9
Convolutional Neural Network (CNN)	7
Naive Bayes	1

Table 2. Detailed detection of environmental parameters.

Environmental Parameters Detected	Number of frequencies
Heavy Metals including (copper (Cu), arsenic (As), cadmium (Cd), chromium (Cr), nickel (Ni), lead (Pb), zinc (Zn), iron (Fe))	47
Soil Characteristics including (soil organic carbon, soil organic matter (SOM), surface soil moisture, soil loss, soil erosion, soil texture, clay content, soil pH, soil salinity, evaporation, soil fauna, soil microbes, electrical conductivity (EC), calcium carbonate equivalent (CCE))	20
Vegetation Properties including (land cover, leaf area index (LAI), cropland suitability assessment, crop productivity, tree counting)	13
Plastic Pollution including (microplastics pollution, plastics polyethylene (PE), polypropylene (PP), polyvinyl chloride (PVC), polyethylene terephthalate (PET), polystyrene (PS), acrylonitrile butadiene styrene (ABS), ethylene vinyl acetate (EVA), polyamide (PA), polycarbonate (PC), polymethyl methacrylate (PMMA))	11
Transition Metals including (Ti (titanium), V (vanadium), Cr (chromium), Mn (manganese), Fe (iron), Co (cobalt), Cu (copper), Zr (zirconium), Nb (niobium), Mo (molybdenum), Cd (cadmium), Hf (hafnium), W (tungsten), M (molybdenum))	11
Alkali and Alkaline Earth Metals including (Na (sodium), Mg (magnesium), K (potassium), Ca (calcium), Sr (strontium), Ba (barium), Cs (cesium))	8
Lanthanides or Rare Earth Elements including (Ce (cerium), Pr (praseodymium), Nd (neodymium), Y (yttrium), La (lanthanum))	5
Nonmetals including (Si (silicon), P (phosphorus), S (sulfur), Br (bromine))	4
Post-Transition Metals including (aluminum (Al), gallium (Ga))	3
Actinides including (Th (thorium), U (uranium))	3
Other pollutants (oil spill, dust and its sources, pollution by urban influence on Inland Marsh)	3
Multi-mycotoxin contamination (such as deoxynivalenol and zearalenone)	2

 doi:10.20944/preprints202503.0068.v1

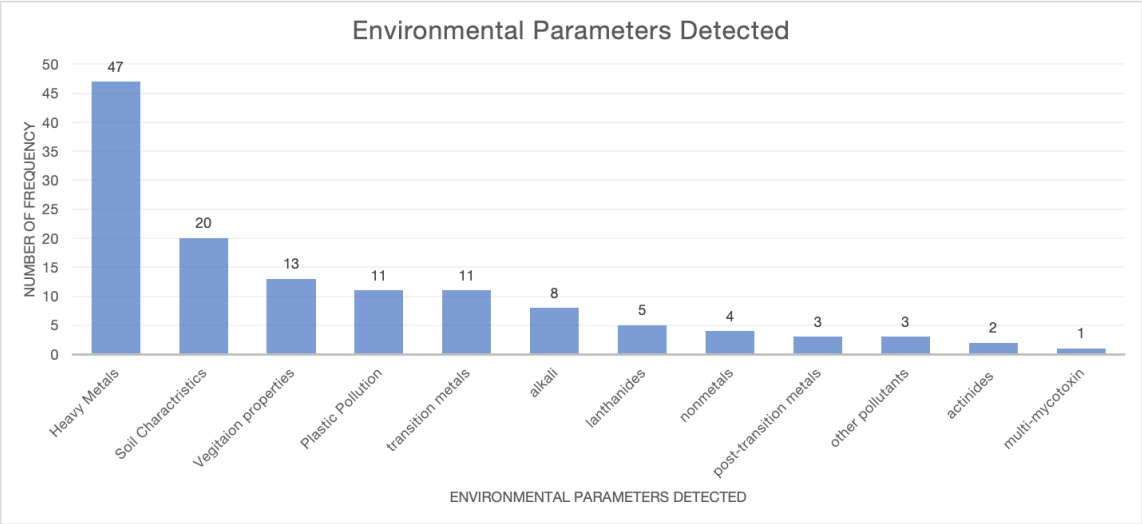

Figure 5. Environmental parameters studied in the selected works.

The number of samples used for validation is crucial in machine learning applications, as a higher number of samples enhances the accuracy of data extraction from satellite imagery. However, the specific number of samples required can vary depending on factors such as task complexity, data diversity, and model architecture. To prevent overfitting and reduce errors, an appropriate range of samples is typically defined during the programming phase. Among the selected articles, 34 studies utilized soil samples for validation [31,32,34,35,37–48,52–59,63–67,69,71–73,75,79].

The number of samples varied significantly across studies, ranging from as few as 6 cases to as many as 15,188 cases. This variation is influenced by factors such as the geographic extent of the study area, data accessibility, and sampling feasibility. Among the 28 articles mentioned, 24 studies evaluated more than one hundred samples, indicating a robust validation process [14,31,34,35,37–40,44,45,47,48,54–57,59,65,75].

Given that satellite imagery primarily detects surface soils, the sampling depth is another critical factor in soil contaminant studies. Sampling depth provides essential insights into the distribution of soil contaminants. In the reviewed articles, 13 studies specified sampling depths, ranging from 5 to 30 cm [31,37,45,53,58,59,63,67,70,71,75,79,81], with the majority focusing on depths between 20 and 30 cm (8 out of 13 studies) [31,45,59,63,71,75,79,81]. This depth range is significant as it captures the root zone of most crops, which is the area most affected by surface contamination.

Sampling depth is important for several reasons. Surface soils (0-5 cm) are often directly exposed to pollutants, making them reliable indicators of recent contamination events. In contrast, deeper sampling (5-30 cm) offers insights into the vertical distribution and potential leaching of contaminants, which is essential for assessing long-term environmental impact and understanding pollutant behavior within the soil profile [53].

Additionally, the root zone, typically within the 20-30 cm range, is where most plant roots are concentrated. Contaminants in this zone can directly affect plant health and crop yield, underscoring the importance of monitoring this depth for agricultural management. Analyzing soil samples from these depths allows researchers to better predict contaminant uptake by plants, which is crucial for food safety and public health [59].

Moreover, sampling at varied depths can reveal the movement and transformation of contaminants through soil layers, influenced by factors such as soil texture, organic matter content, and water flow. This information is essential for developing effective remediation strategies and implementing sustainable land-use practices. Depth-specific sampling improves understanding of soil contamination processes and supports more accurate environmental risk assessments.

Most results obtained from the reviewed articles on soil characteristic identification exhibited acceptable levels of accuracy. For example, Agrawal et al. [58] demonstrated strong evidence that soil

 doi:10.20944/preprints202503.0068.v1

arsenic contamination can be detected using Hyperion satellite hyperspectral data when combined with preprocessing and machine learning. Similarly, Azizi et al. [51] highlighted the effectiveness of machine learning methods in utilizing readily available environmental data to predict the presence of heavy metals on a large scale. These predictions have significant implications for sustainable management decision-making, particularly in agriculture and environmental monitoring.

Overall, this systematic review has provided valuable insights into the potential and effectiveness of satellite-based soil contaminant detection using machine learning methodologies. The analysis of different satellites, machine learning techniques, evaluation metrics, and environmental parameters highlights both the strengths and limitations of current approaches. These findings contribute meaningfully to understanding soil pollution monitoring and management strategies, paving the way for advancements in this important field of study.

4. Conclusions

In conclusion, this systematic review explored the detection of soil pollution using satellite imagery and machine learning methods, following the PRISMA methodology. Sentinel-2 and Landsat 8 were identified as the most frequently employed satellites for soil pollution detection, proving effective in various domains, including land-use mapping, environmental monitoring, and natural resource management.

The review found that Sentinel-2 was used in seven instances to detect soil properties and in four instances to detect heavy metals. Similarly, Landsat 8 was used in seven instances for detecting both heavy metals and soil properties. The high spatial resolution and frequent global coverage provided by these satellites have proven invaluable for soil pollution monitoring.

Machine learning methods played a crucial role, with Random Forest (RF) emerging as the most prevalent method, used in 33 out of 47 cases. RF's ability to handle large datasets with high-dimensional features, along with its robustness against noise and outliers, was particularly advantageous. Other notable methods included Support Vector Machine (SVM), Cubist, Artificial Neural Network (ANN), and Partial Least Squares Regression (PLSR), each contributing to classification and prediction tasks in soil contaminant detection.

The review emphasized the importance of sample size and sampling methods, with most studies involving over one hundred samples. Sample sizes ranged from 6 to 15,188 cases, influenced by factors such as site size and data accessibility. Depth-specific sampling, ranging from 5 to 30 cm, was critical for capturing contamination profiles, with the majority of studies focusing on a depth of 20 to 30 cm—significant for capturing the root zone of most crops and areas most impacted by surface contamination.

Performance indicators such as Root Mean Square Error (RMSE) and the coefficient of determination (R^2) were frequently used, appearing in 30 and 25 studies, respectively. These metrics provided a comprehensive assessment of model performance, highlighting the accuracy and reliability of soil contaminant detection methods.

Integrating multiple satellite datasets with advanced machine learning techniques holds significant potential for enhancing soil pollution monitoring and management. By combining data from various satellite sources, researchers can address limitations of individual satellite systems, thereby improving the overall accuracy and depth of analysis.

In summary, satellite imagery for soil contaminant detection is a promising approach with growing validation in the literature. The extensive use of machine learning models underscores the evolving nature of this research field. As technology advances and more sophisticated satellites are launched, the precision and applicability of these methods are expected to improve, paving the way for more effective environmental conservation strategies.

Future research will particularly focus on addressing critical gaps, such as tracking the origin and evolution of pollution sources over time. A key area for exploration involves the temporal aspect, requiring the acquisition of satellite images over months or years to track changes and identify emerging pollution sources. While existing research has made strides in this direction, deeper

 doi:10.20944/preprints202503.0068.v1

integration of advanced machine learning methods, diverse satellite technologies, and additional parameters is necessary to achieve comprehensive soil pollution assessment.

Funding: This work was financially supported by Base Funding - UIDB/04028/2020 and Programmatic Funding - UIDP/04028/2020 of the Research Center for Natural Resources and Environment – CERENA - funded by national funds through the FCT/MCTES (PIDDAC). Furthermore, heartfelt gratitude is extended for the research grant accompanying this thesis. The grant was awarded under the Statute of Research Fellows amended by Deliberation n.° 1301/2021, published in D.R., 2nd series, on 22 December 2021, have played a vital role in facilitating the completion of this study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the data presented in this study are not publicly available in compliance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology employed in this research. PRISMA guidelines emphasize the necessity of safeguarding the integrity and quality of systematic reviews and meta-analyses by maintaining controlled access to data. To ensure accurate and responsible data interpretation, our study follows PRISMA's recommendations by making the data available exclusively upon request from the corresponding author. This approach allows us to uphold the rigorous standards set by PRISMA while also fostering transparency and scientific collaboration by providing access to interested parties who can ensure the proper use and interpretation of our data.

Conflicts of Interest: The authors declare no conflict of interest.

 doi:10.20944/preprints202503.0068.v1

Appendix A

<table>
<tr><th>Ref.</th><th colspan="3">Study Characteristics</th><th colspan="2">Validation</th><th colspan="2">Study Results
performance</th><th rowspan="2">Result</th></tr>
<tr><th>Ref.</th><th>Satellite Name</th><th>Method of ML</th><th>Environmental Parameters Detected</th><th>Validation Type</th><th>Validation Sample numbers</th><th>Performance Metrics</th><th>Best Performance Results</th></tr>
<tr><td>[31]</td><td>Landsat 4 (TM), Landsat 5 (TM), Landsat 6 (ETM), Landsat 7 (ETM+), Landsat 8, Sentinel 2</td><td>Random Forest (RF)</td><td>Cr, Fe, Ni, and Zn</td><td>soil samples using an auger (0–20 cm depth).</td><td>360 soil samples.</td><td>(R2), (MAE), (MSE).</td><td>
<table>
<tr><th>Target</th><th>MAE(Cal)</th><th>R2(Cal)</th><th>MAE (Val)</th><th>R2(Val)</th></tr>
<tr><td>Cr</td><td>9.18 ± 1.19</td><td>0.16</td><td>8.90 ± 0.40</td><td>0.23</td></tr>
<tr><td>Fe</td><td>111.34 ± 24.35</td><td>0.55</td><td>78.77 ± 30.11</td><td>0.61</td></tr>
<tr><td>Ni</td><td>4.74 ± 0.47</td><td>0.13</td><td>3.33 ± 0.38</td><td>0.16</td></tr>
<tr><td>Zn</td><td>11.00 ± 2.15</td><td>0.22</td><td>8.44 ± 1.20</td><td>0.20</td></tr>
</table>
</td><td>The clearest discrimination of soil PTEs was obtained from SYSI using a long-term Landsat 5 collection over 35 years. Satellite data could efficiently detect the contents of PTEs in soils due to their relation with soil attributes and parent materials.</td></tr>
</table>

 doi:10.20944/preprints202503.0068.v1

[32]	Terra, Aqua, NOAA satellites, Landsat satellites, PROBA-V	the general regression neural network (GRNN), long short-term memory (LSTM), gated recurrent unit (GRU), and Bidirectional LSTM (Bi-LSTM)	Leaf area index (LAI)	Reference maps were collected from 2000 to 2016 at 47 sites from Bigfoot from VALERI and ImagineS networks with different dominant biome types.	79 available high-resolution LAI reference maps.	number of samples points (N), R2, RMSE, bias, and the percentage of pixels meeting the target accuracy requirement (P)	The results show that GLASS V6 LAI achieves higher accuracy, with a root mean square (RMSE) of 0.92 at 250 m and 0.86 at 500 m, while the RMSE is 0.98 for PROBA-V at 300 m, 1.08 for GLASS V5, and 0.95 for MODIS C6 both at 500 m.	GLASS V6 LAI product is more spatiotemporally continuous and has higher quality in terms of presenting more realistic temporal LAI dynamics when the surface reflectance is absent for a long period owing to persistent cloud/aerosol contaminations. The results indicate that the new Bi-LSTM deep learning model runs significantly faster than the GLASS V5 algorithm, avoids the reconstruction of surface reflectance data, and is resistant to the noises (cloud and snow contamination) or missing values

 doi:10.20944/preprints202503.0068.v1

								contained in surface reflectance than other methods, as the Bi-LSTM can effectively extract information across the entire time series of surface reflectance rather than a single time point.
[33]	Landsat 8	random forest (RF)	multi-mycotoxin contamination (such as deoxynivalenol and zearalenone)	prediction results were validated with the Dutch data in the testing set. The model was then run with the input variables of the external validation set. The predicted model results for 2019 and 2020 were compared with the analyzed mycotoxin data (per contamination level) in these two years, separately.	-	Confusion metrics, accuracy, and generalization ability	internal and external validation resulted in 0.90–0.99 prediction accuracy.	It can be concluded that the use of machine learning algorithms for mycotoxin prediction in risk levels at the regional level in Europe provides good prediction results. Such models can be used by collectors, traders, and food safety authorities for logistics in the wheat supply chain, improved

 doi:10.20944/preprints202503.0068.v1

								mycotoxin control, and risk-based testing.
[34]	Landsat 8	support vector regression (SVR), partial least square regression (PLSR) and artificial neural network (ANN)	soil copper (Cu)	soil samples were collected in this study area in 2015, and the Cu concentrations of samples were analyzed and recorded.	138 soil samples with lab-measured Cu concentrations.	coefficient of determination (R2), root mean square error (RMSE), mean absolute error (MAE) and standard error (SE)	The mean adjusted R2 obtained by SVR using 20 repeated 6-fold cross-validations on 138 soil samples increases from 0.433 to 0.641. The mean R2 of PLSR and ANN increase from 0.568 to 0.618 and from 0.476 to 0.528 separately, indicating the necessity and benefit of feature extraction and selection. Although ANN is a popular regression method, in our work, SVR outperforms ANN by achieving a mean R2 of 0.641, which is 21.4% higher than ANN. RMSE, MAE and SE also support the highest generalization capability of SVR.	The preferred model with the highest R2 obtained by SVR is selected to estimate the Cu concentration in soil over the study area. Compared to the interpolation map, the Cu concentration distribution map generated by the recommended pipeline gives the pixel-based Cu estimation with more spatial detail and wider spatial coverage. It also shows a consistent spatial pattern with the ground-truth

 doi:10.20944/preprints202503.0068.v1

								land cover classification map. The results show this model's ability to perform large-scale soil (HMC heavy metal contamination) mapping from widely available satellite imagery.
[35]	Sentinel 1, Sentinel 2	Random Forest (RF)	oil spill	Sample locations of oil-free sites that were not located in close proximity to the observed oil spill sites were selected. A buffer zone of 500 m was implemented around the spill areas to exclude all existing spill points from the in-situ observations.	n = 553 each candidate sample point was classified into one of six thematic categories based on expert knowledge resulting from a number of on-site visits in 2019 and 2020.	Overall accuracy, Kappa, LI 95 %, UI 95 %, F1 scores, The user's classification accuracy (UA), producer's classification accuracy (PA).	lowest overall accuracies for Oil spill-I and Oil spill-II were 91.4 % and 85.0 %, respectively.	The mapping of terrestrial oil spills with freely available Sentinel satellite images may thus represent an accurate and efficient means for the regular monitoring of oil-impacted areas. Such tools can be used to create an open access database for oil mapping, which would enable indigenous communities to document oil

 doi:10.20944/preprints202503.0068.v1

								pollution from the remote areas they inhabit and provide local communities, journalists, and civil society organizations with reliable proof of environmental damage.
[36]	Landsat 8	Pixel-based RNN system (Pix RNN), Pixel-based single-image NN system (Pix single), Pixel-based multi-image NN system (Pix multi), Patch-based single-image NN system (Patch Single), Patch-based multi-image NN system, Proposed patch-based RNN (PB-RNN)	Land cover classification	Obtained ancillary data from the Florida Cooperative Land Cover Map first and performed corrections by comparing it with GPS guided field observations and the high-resolution images from Google Earth.	A series of 23 Landsat 8 images were used in this study, to evaluate the proposed method on a test site within the Florida Everglades Ecosystem.	Overall accuracy (OA), Overall kappa(kappa) , Error Matrix.	The proposed system achieves 97.21% classification accuracy while the pixel-based single-image NN system achieves 64.74%. the proposed system achieves 0.97 overall kappa while the pixel-based single-image NN (Neural Network) system achieves 0.58.	The classification results show that the proposed system achieves significant improvements in both the overall and categorical classification accuracy.

 doi:10.20944/preprints202503.0068.v1

<table>
<tr>
<td>[37]</td>
<td>Space Shuttle Endeavour (SRTM-1)</td>
<td>Random Forests (RF), Cubist, Linear Model, Support Vector Machine, K Nearest Neighbor (KNN)</td>
<td>Pb, Zn, Ba, Fe, Al, and Cr.</td>
<td>Collected at a depth of 0–10 cm.</td>
<td>120 soil samples.</td>
<td>MAE, RMSE, R2.</td>
<td>
<table>
<tr><th>Target</th><th>Best Method</th><th>MAE(mg kg−1)</th><th>RMSE(mg kg−1)</th><th>R2(mg kg−1)</th></tr>
<tr><td>Lead</td><td>Cubist</td><td>120.97</td><td>264.03</td><td>0.795</td></tr>
<tr><td>Zinc</td><td>Cubist</td><td>76.33</td><td>193.11</td><td>0.801</td></tr>
<tr><td>Barium</td><td>Cubist</td><td>21.24</td><td>30.820</td><td>0.55</td></tr>
<tr><td>Chrome</td><td>Cubist</td><td>5.31</td><td>7.21</td><td>0.37</td></tr>
<tr><td>Iron</td><td>Cubist</td><td>5179.49</td><td>9357.34</td><td>0.90</td></tr>
<tr><td>Aluminum</td><td>SVM</td><td>2089.88</td><td>2809.74</td><td>0.84</td></tr>
</table>
</td>
<td>In general, the Cubist algorithm produced better results in predicting the contents of Pb, Zn, Ba and Fe compared to the other tested models. For the Al contents, the Support Vector Machine produced the best prediction. The methodology structure reported in this study represents an alternative for fast, low-cost prediction of PTEs in soils, in addition to being efficient and economical for monitoring potentially contaminated areas and obtaining quality reference values for soils.</td>
</tr>
</table>

 doi:10.20944/preprints202503.0068.v1

[38]	PROBA-V	Random forest (RF), support vector machine (SVM)	Cropland Suitability Assessment.	A total of 119 covariates were used per the individual prediction of yearly cropland suitability classes for soybean cultivation, consisting of 47 climates, 24 soil, 6 topographic and 42 vegetation covariates.	Samples, with a total of 119 covariates being utilized per yearly suitability assessment.	Accuracy assessment, R2, RMSE	Random forest (RF) mean overall accuracy of 76.6% to 68.1% for Subset A and 80.6% to 79.5% for Subset B.	RF produced superior suitability assessment results to SVM in cases of moderate sample count and a high amount of complex input covariates. The proposed method overcomes the limitations of the conventionally used GIS-based multicriteria analysis, and could turn the attention to machine learning in future cropland suitability determination studies.
[39]	Sentinel 2	support vector machine (SVM), random forest (RF), Classification and Regression Trees (CART)	Land Cover	Total data included: Corn crop: 113 Sorghum crop: 547 Water bodies: 190 Land in recovery: 226	For the test of the classified maps, 30% of the sample points were used: 742 for the spring–summer season and 868 for the	Overall accuracy (OA), Kappa index (KI)	The results in overall accuracy were: 0.99% for the support vector machine, 0.95% for the random forest, and 0.92% for classification and regression trees. The kappa index was: 0.99% for the support vector machine, 0.97% for the random forest, and 0.94% for classification and regression trees.	The area and seasons studied presented a high rate of humidity, which made the research difficult. On the other hand, the execution

 doi:10.20944/preprints202503.0068.v1

				Urban areas: 66 Sandy areas: 117 Tropical rainforest: 237 Others: 170	autumn–winter season			capacity of the Google Earth Engine platform proved to be effective in land-use analysis and classification. The methods used for land-use classification and crops of sorghum and corn were SVM, RF, and CART, which obtained different results.
[40]	Sentinel 2, Landsat 8	partial least square regression (PLSR), extreme learning machine (ELM).	soil organic carbon	Two soil sampling campaigns (50 soil samples on October 2015 and 145 soil samples on March 2016) were operated to collect the surface soil samples (0–15 cm) using a grid soil sampling strategy with 130 m.	195 surface soil samples were collected	(RMSE), R2, ratio of performance to interquartile range (RPIQ)	Hyperspectral images were successfully used to predict the SOC stock, SOC, and SBD through PLSR and ELM, while ELM (RPIQ = 2.03, 1.97, 1.64) outperformed PLSR (RPIQ = 1.83, 1.97, 1.53); Sentinel 2 images and ELM obtained the best prediction results (RPIQ = 1.45, 1.25, 1.26);	This study further confirmed the good prediction abilities of the time-series multispectral remote sensing images in low relief farmland regions. Lastly, this mapping strategy can provide additional valuable information for agricultural

 doi:10.20944/preprints202503.0068.v1

								management and carbon cycle.
[41]	GAOFEN 1	Neu-SICR algorithm	Surface soil moisture	In situ soil moisture values observed by probes at soil moisture observatories among the soil climate analysis network (SCAN) were traced and adopted.	11 soil moisture observatories from the soil climate analysis network (SCAN) could be accessed.	Average relative error (ARE), the universal image quality index (UIQI)	ARE: 13.18 and UIQI: 0.3143.	The new algorithm enhances the temporal resolution of high spatial resolution remote sensing regional soil moisture observations with good quality and can benefit multiple soil moisture-based applications and research.
[42]	GAOFEN 5	random forest (RF), ExtraTrees (ET), Adaptive Boosting (ADB), Gradient Descent Boosting Trees (GDB), eXtreme Gradient	Arsenic	Systematic grid sampling was conducted, and sampling locations were set based on a 40mregular grid. In each sample location, the soil sample was filled with 250 ml wide-	In the whole study area, a total of 976 topsoil samples (0–30 cm) were collected.	r, RMSE, MAE	RF also maintained a relatively higher level of accuracy (r = 0.56) when the sampling grids increased to 100 m, which was higher than that of GIMs under a 50 m sampling grid (r = 0.42).	This study demonstrates that machine learning based on satellite visible and near-infrared reflectance spectroscopy (VNIR) is a promising approach to map soil arsenic

 doi:10.20944/preprints202503.0068.v1

		Boosting (XGB)		mouth sampling bottles, the sample locations were confirmed by real-time kinematic (RTK) mobile station positioning technology.				contamination at brownfield sites with high accuracy and low cost. The RF method was found to render the best performance (r = 0.78), reducing 30 % of prediction errors compared with traditional GIMs.
[43]	Worldview 2	U-net convolutional neural network (CNN)	Microplastics pollution	Surface soil (2 cm) were randomly collected from the selected area (50 cm * 50 cm) using a stainless shovel for each subsample.	6 mixed samples were collected at each site, each mixed sample being composed of three subsamples, included mulching soil and non-mulching soil.	-	-	The results revealed that the abundance of MPs in soil mulched by dust-proof nets ranged from 272 to 13,752 items/kg. Large-sized particles (>1000 mm) made up a significant proportion (49.83%) of MPs in the study area. This study will highlight the understanding of soil MPs pollution and its potential

 doi:10.20944/preprints202503.0068.v1

								environmental impacts for scientists and policymakers. It provides suggestions for decisionmakers to formulate effective legislation and policies, so as to protect human health and protect the soil and the wider environment.
[44]	Landsat 4, Landsat 5 (TM)	Random Forest (RF), Extreme Gradient Boosting (XGBoost)	Soil pH	The full set contains soil profiles with the descriptions of geographic location, genetic horizon thickness, organic matter, pH, texture (particle-size distribution), total nitrogen, total phosphorus and bulk density. The pH was measured	4700 soil profiles were available from China's Second National Soil Inventory.	coefficient of determination (R2), the root mean squared error (RMSE) and Lins's Concordance Correlation Coefficient (CC)	The combined two models' root mean squared error (RMSE) was an acceptable 0.71pH units per point, and Lin's Concordance Correlation Coefficient was 0.84.	This map can provide a benchmark against which to evaluate the impacts of changes in land use and climate on the soil's pH, and it can guide advisors and agencies who make decisions on remediation and prevention of soil acidification, salinization and pollution by heavy metals, for which we

 doi:10.20944/preprints202503.0068.v1

				with a pH meter in a suspension of soil in water with a soil: water ratio of 1:2.5.				provide examples for cadmium and mercury.
[45]	SPOT 5 (Satellite Pour l'Observation de la Terre)	Random forest (RF), geographically weighted regression (GWR)	zinc (Zn)	A sampling site was then randomly chosen in the grids during the sampling process. The geographical coordinates of the sampling sites were recorded using a GPS (global positioning system) receiver. The samples were collected from vegetated or exposed soils in parks, gardens, greenbelts, etc., and impervious areas were	221 soil samples	Accuracy, R values, R2, RMSE.	The RF and GWR models were established using the key environmental covariates, with leave-one-out cross-validated R values of 0.68 and 0.58 and Root Mean Square Errors of 0.51 and 0.57, respectively.	The results showed that urban functional type, geology, NDVI, elevation, slope, and aspect were key environmental covariates. Compared with land use types, urban functional types could better reflect the spatial variation in Zn.

 doi:10.20944/preprints202503.0068.v1

<table>
<tr><td></td><td></td><td></td><td></td><td>avoided. At each site, we collected approximately 1.5 kg soil samples (0–20 cm) using a shovel, from which plant residues and artificial deposits were removed.</td><td></td><td></td><td></td><td></td></tr>
<tr><td>[46]</td><td>ZY-1-02D satellite</td><td>CART, MLP, SVM, Gaussian process regression (GPR), K-nearest neighbor (KNN), kernel ridge regression (KRR), AdaBoost.</td><td>Heavy metal (Cr, Cu, and As)</td><td>Based on remote sensing images, the distribution of farmland in the study area was determined, and the sampling points were set at one-kilometer intervals. Through field investigation, we adjusted the preset locations of sampling points and the sequence and route of sample collection</td><td>81 soil samples.</td><td>R2, RMSE, RPD</td><td>For Cr, Cu, and As, the determination coefficients (R2) of the verification set were 0.66, 0.61, and 0.74, respectively for the AdaBoost model.
<table>
<tr><th>Target</th><th>R^2_c</th><th>$EMSE_c$</th><th>R^2_p</th><th>$EMSE_p$</th><th>RPD</th></tr>
<tr><td>Cr</td><td>0.73</td><td>3.71</td><td>0.66</td><td>4.52</td><td>2.06</td></tr>
<tr><td>Cu</td><td>0.69</td><td>1.94</td><td>0.61</td><td>2.36</td><td>1.85</td></tr>
<tr><td>As</td><td>0.87</td><td>0.73</td><td>0.74</td><td>0.95</td><td>1.72</td></tr>
</table></td><td>In summary, the Stacked AdaBoost ensemble learning model provides detailed and reliable data for agricultural ecological protection and industrial pollution control, allowing the effective management of heavy metal pollution sources.</td></tr>
</table>

 doi:10.20944/preprints202503.0068.v1

<table>
<tr>
<td>[47]</td>
<td>Sentinel 2A</td>
<td>Partial least squares regression (PLSR), backward propagation neural network (BPNN), Random Forest (RF)</td>
<td>The Cd, Pb, soil organic matter (SOM), pH, and Fe</td>
<td>A portion of the soil sample passing through a 100-mesh nylon sieve was used to determine the Cd, Pb, and Fe contents. Another part of the soil sample was passed through a 10-mesh nylon sieve to determine the SOM content and pH value. Cd and Pb contents were measured by inductively coupled plasma–mass spectrometry.</td>
<td>640 samples from the surface soils</td>
<td>R2
NRMSE
RPD</td>
<td>Relatively satisfactory estimates of Cd and Pb contents in farmland of the study area (maximum R2val (determination coefficient of the validation set) = 0.60 for Cd and R2val = 0.63 for Pb) were obtained.
<table>
<tr><th>Target</th><th>Best Method</th><th>R2</th><th>NRMSE</th><th>RPD</th></tr>
<tr><td>Cd (Original images)</td><td>RF</td><td>0.46</td><td>0.101</td><td>1.74</td></tr>
<tr><td>Cd (Unmixed images)</td><td>RF</td><td>0.50</td><td>0.098</td><td>1.80</td></tr>
<tr><td>Pb (Original images)</td><td>RF</td><td>0.52</td><td>0.066</td><td>1.82</td></tr>
<tr><td>Pb (Unmixed images)</td><td>RF</td><td>0.57</td><td>0.062</td><td>1.94</td></tr>
<tr><td>Cd (Original images) Double data images</td><td>RF</td><td>0.55</td><td>0.093</td><td>1.89</td></tr>
<tr><td>Cd (Unmixed images)</td><td>RF</td><td>0.60</td><td>0.088</td><td>2.01</td></tr>
<tr><td>Pb (Original images) Double data images</td><td>RF</td><td>0.60</td><td>0.060</td><td>2.01</td></tr>
<tr><td>Pb (Unmixed images) Double</td><td>RF</td><td>0.63</td><td>0.057</td><td>2.10</td></tr>
</table>
</td>
<td>The results of the study provide a theoretical basis and methodological reference for the rapid prediction of Cd and Pb contents in regional farmland.</td>
</tr>
</table>

 doi:10.20944/preprints202503.0068.v1

							data images	
[48]	Gaofen 5, PRISMA	convolutional neural network (CNN), random forest (RF) and support vector machine (SVM)	Plastics polyethylene (PE), polypropylene (PP), polyvinyl chloride (PVC), polyethylene terephthalate (PET) and polystyrene (PS), some important varieties of industrial plastics types such as acrylonitrile butadiene styrene (ABS), ethylene vinyl acetate (EVA), polyamide (PA), polycarbonate (PC), and polymethyl methacrylate (PMMA).	Different samples with varying optical properties (color, brightness, transmissivity) have been selected for each plastic type.	Over 3000 samples were collected within the three formers mentioned spectral libraries.	Recall, precision, F1-score, overall accuracy (OA), Kappa.	The performance of the three (Satellite, airborne and laboratory) models is roughly balanced for the validation of the spectral data with an overall accuracy of 97%, 96%, and 95% for the CNN, RF, and SVM, models respectively. In principle, it can be stated that the RF classifier produced very good and reliable results for the data of both sensors.	The RF was used to classify the ten types of plastics in GF-5 and PRISMA satellite recordings of the same area. In comparison of both sensor systems, the RF produced high quality and transferable results for detecting plastic mainly related to greenhouses, sport fields, photovoltaic constructions and industrial sites.
[49]	Terra, Aqua	Apriori algorithm	dust	-	-	-	The accuracy of the identified SDSs was estimated at 83.7% using the verification points.	The results revealed that Apriori's ability to provide

 doi:10.20944/preprints202503.0068.v1

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>generalizable association rules is a robust algorithm for Data-Driven Soil Mapping (DSSM).</td></tr>
<tr><td>[50]</td><td>Sentinel 2</td><td>Random Forest (RF)</td><td>Land cover classification</td><td>Different NBS actions were simulated based on the reforestation of specific areas and were compared with the actual situation.</td><td>-</td><td>Error matrix, The Error of Commission (EC), Error of Omission (EO), Producer Accuracy (PA), User Accuracy (UA).</td><td>Land cover change in the Umia Basin, was successfully mapped with Sentinel-2 images with an OA of 77% and 83%.</td><td>It has been proven that the use of reforestation upstream only is almost as beneficial as reforestation in the entire catchment and is economically more viable. This confirms that the methodology used reduces flood hazard, despite the territorial complexity, facilitating decision making on the use of NBS.</td></tr>
<tr><td>[51]</td><td>Landsat 8</td><td>Random forest (RF), and Cubist.</td><td>Heavy metals (Ni, Fe, Cu, Mn)</td><td>Surface layers (0–20 cm depth).</td><td>346 soil samples</td><td>R2, RMSE</td><td><table><tr><th>Target</th><th>Best Method</th><th>EMSE</th><th>R^2</th></tr><tr><td>Fe</td><td>RF, cubist</td><td>0.20</td><td>0.73</td></tr><tr><td>Mn</td><td>Cubist</td><td>0.19</td><td>0.55</td></tr><tr><td>Cu</td><td>RF</td><td>0.19</td><td>0.60</td></tr><tr><td>Ni</td><td>RF</td><td>0.15</td><td>0.67</td></tr></table></td><td>This study proved the high capability of machine learning methods to use easily available environmental</td></tr>
</table>

 doi:10.20944/preprints202503.0068.v1

								data to predict studied heavy metals in the large scale that are essential for decision making in sustainable management in agricultural and environmental concerns.
[52]	Sentinel 2, CRYOSAT 2, Jason 1.	LSTM model (deep learning), RNN, CNN.	Heavy metal pollution (copper).	Selected 19 different types of data, including basic geological data and anomaly data, as experimental datasets.	As the 19 types of data are input into the stacked LSTM model.	Overall accuracy (OA).	The validation dataset includes a total of 31 copper mineral occurrences, of which 9 are classified as grade IV and 17 are classified as grade V, accounting for 83.87% of all mineral occurrences.	Using the optimized stacked LSTM model to integrate multisource geological features and mine the internal rules of feature information has a positive effect on improving the risk assessment of heavy metal pollution.
[53]	Landsat 8	Random Forest (RF), generalised boosting methods (GBM), generalised	Origin of trace metals (Na, Mg, Al, Si, P, S, K, Ca, Ti, V, Cr, M, Fe, Co, Ni, Cu, Zn, Ga, As, Br, Rb, Sr,	Soil samples were collected from the topsoil (0–5 cm). The sampling depth of 0 to 5 cm guarantees	79 soil samples having different Nemerow index values were considered	Receiver Operating Characteristic (ROC), Area Under the Curve (AUC), And OA.	RF had the best performance with an accuracy of 83%. The evaluation of polluted soil areas showed that the landforms ‘steep hills’ and ‘valley’ contributed the most with 51%and 27%in the riparian zone, respectively. The landform ‘plain’ had the highest contribution (28%) in sediment yield with a GOF of 0.72 in early-winter events.	Overall, the new proposed approach enables to better trace the origin of suspended sediments and

 doi:10.20944/preprints202503.0068.v1

<table>
<tr><td></td><td></td><td>linear models (GLM)</td><td>Y, zr, Nb, Mo, Cd, Cs, Ba, La, Ce, Pr, Nd, Hf, W, Pb, Th, U, Soil Organic Carbon)</td><td>that only surface materials that potentially may be displaced by surface runoff are sampled.</td><td>for spatial modelling.</td><td></td><td></td><td>trace elements discharge into the river environment.</td></tr>
<tr><td>[54]</td><td>Landsat 7, Landsat 8</td><td>Cubist, random forest (RF).</td><td>Toxic elements (PTEs) (As, Cr, Cu, Ni, Pb and Zn) and modified pollution index (MPI).</td><td>To have an accurate estimation, a stratified simple random sampling method based on a grid of 400 ha was chosen. Since there was no heterogeneity in landform, geology, vegetation, land management, etc., the grid lines were assumed as the strata boundaries</td><td>129 surficial soil samples.</td><td>r, RMSE, bias, CCC, Enrichment factors (EFs) of PTEs and the Modified Pollution Index (MPI).</td><td>Calibration dataset (in the bag) R2 for all the heavy metals between 0.84-0.8, Cubist 0.19-0.45.
<table><tr><th>Target</th><th>Best Method</th><th>R2</th><th>Concordance</th><th>NRMSE</th><th>Bias</th></tr><tr><td>AS</td><td>RF</td><td>0.25</td><td>0.30</td><td>0.019</td><td>0.03</td></tr><tr><td>Cr</td><td>RF</td><td>0.23</td><td>0.27</td><td>0.002</td><td>0.00</td></tr><tr><td>Cu</td><td>RF</td><td>0.20</td><td>0.39</td><td>0.014</td><td>0.04</td></tr><tr><td>Ni</td><td>RF</td><td>0.21</td><td>0.25</td><td>0.004</td><td>0.00</td></tr><tr><td>Pb</td><td>RF</td><td>0.28</td><td>0.31</td><td>0.006</td><td>0.02</td></tr><tr><td>Zn</td><td>RF</td><td>0.23</td><td>0.28</td><td>0.002</td><td>0.00</td></tr><tr><td>MPI</td><td>RF</td><td>0.27</td><td>0.37</td><td>0.014</td><td>0.00</td></tr></table></td><td>The results showed that Random Forests performed well in estimating EFs of several PTEs. Spectral indices using NIR and SWIR bands were key to predict these PTEs and MPI. The digital maps demonstrated that the study area was enriched with As, Cu and Pb at moderate to significant levels.</td></tr>
<tr><td>[55]</td><td>Gaofen 5</td><td>Random Forest (RF), the extreme learning machine (ELM),</td><td>Soil heavy metals (Zn, Ni, and Cu)</td><td>The sampling route was arranged according to FOREGS Geochemical</td><td>110 topsoil samples</td><td>R2, RMSE, MAE</td><td>The estimation accuracy was significantly improved by using the Decision Stump algorithm.
<table><tr><th>Target</th><th>Best Method</th><th>R2</th><th>RMSE (Mg kg-1)</th><th>MAE (Mg kg-1)</th></tr><tr><td>Zn</td><td>RF</td><td>0.77</td><td>9.54</td><td>7.39</td></tr></table></td><td>This paper revealed that the GF-5 can be one of the reliable satellite’s hyperspectral</td></tr>
</table>

 doi:10.20944/preprints202503.0068.v1

		the support vector machine (SVM), the back-propagation neural network (BPNN)		Mapping Field Manual			<table><tr><td>Ni</td><td>RF</td><td>0.62</td><td>3.39</td><td>2.56</td></tr><tr><td>Cu</td><td>ELM</td><td>0.56</td><td>5.02</td><td>3.73</td></tr></table>	imageries for mapping soil heavy metals
[56]	Sentinel 1A, Alos Palsar I (SAR), Sentinel 2A	Classification and regression tree (CART), artificial neural network (ANN), random forest (RF), k-nearest neighbors (kNN).	Pollution by urban influenced on Inland Marsh	Samples from fieldwork that took place between 12/01/2018 to 12/04/2018. To collect the samples, two Global Navigation Satellite System (GNSS) Ruide R90-X dual-frequency (L1/L2) receivers were used.	450 samples of the wet meadow.	overall accuracy (OA) producer's accuracy (PA), user's accuracy (UA).	The results showed that the method with the highest overall accuracy was k-NN, with 98.5%. The accuracies for the RF, ANN, and CART methods were 98.3%, 96.0% and 95.5%, respectively. The four classifiers presented accuracies exceeding 95%, showing that all methods have potential for inland marsh delineation.	CART and ANN methods presented the largest variations of the overall accuracy (OA) in relation to the different parameters tested.
[57]	Terra, Aqua (MODIS)	Cubist	Soil loss	In the spatial modelling as well as the performance of the model using the samples not included in the bootstrap, i.e., the out-of-bag (OOB) samples.	100 bootstrap samples to assess the uncertainties	R2, RMSE	Estimate the average erosion rate in Australia to be 4.16 t ha_1 y_1, and the total amount of annual soil loss to be 2788 *106 tones. <table><tr><td>Target</td><td>R2</td><td>RMSE</td></tr><tr><td>Cross Validation Statistics</td><td>0.68</td><td>0.38</td></tr><tr><td>Out of bag statistics</td><td>0.69</td><td>0.02</td></tr><tr><td>Test Set Statistics</td><td>0.71</td><td>0.01</td></tr></table>	Estimation of erosion are generally smaller than previous continental estimates using the Revised Universal Soil Loss Equation (RUSLE), but particularly in

 doi:10.20944/preprints202503.0068.v1

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>croplands, which might indicate that soil conservation practices effectively reduced erosion in Australia.</td></tr>
<tr><td>[58]</td><td>Earth Observing-1 (EO-1)</td><td>Regression Train/Test: (Partial Least Squares Regression (PLSR), Back Propagation Neural Network (BPNN), Random Forest (RF), K-Nearest Neighbors (KNN).) High-Risk Classification : (Support Vector Machine (SVM), Random Forest Classification (RFC), Multi-Layer Perceptron</td><td>Arsenic</td><td>Publicly available data for soil arsenic concentration in the United States between 2005 and 2020 for both aforementioned land covers (in mg/kg) were processed for the top layer of soil (0–5 cm depth).</td><td>A total of 55 bare soil arsenic concentration values, with concentrations between 1.4 mg/kg and 380 mg/kg, were used for regression analysis</td><td>Accuracy, F1-Score, F2-Score, F0.5-Score, Brier Score, R2, RMSE</td><td>
<table>
<tr><th>Target</th><th>R2</th><th>NRMSE</th></tr>
<tr><td>SD + (PLSR)</td><td>0.623</td><td>0.194</td></tr>
<tr><td>SD + (BPNN)</td><td>0.726</td><td>0.144</td></tr>
<tr><td>SD + (RF)</td><td>0.746</td><td>0.136</td></tr>
<tr><td>SD + (KNN)</td><td>0.715</td><td>0.178</td></tr>
<tr><td>GA+SD + (PLSR)</td><td>0.668</td><td>0.162</td></tr>
<tr><td>GA+SD + (BPNN)</td><td>0.704</td><td>0.173</td></tr>
<tr><td>GA+SD + (RF)</td><td>0.805</td><td>0.132</td></tr>
<tr><td>GA+SD + (KNN)</td><td>0.693</td><td>0.171</td></tr>
<tr><td>DA + GA + SD + RF</td><td>0.840</td><td>-</td></tr>
</table>
Comparisons of the evaluation metrics of the three binary classification ML models for the averaged swath data.
<table>
<tr><th>Model</th><th>Accuracy</th><th>F1-Score</th><th>F2-Score</th><th>F0.5-Score</th><th>Brier Score</th></tr>
<tr><td>SVM</td><td>0.647</td><td>0.688</td><td>0.658</td><td>0.751</td><td>0.272</td></tr>
</table>
</td><td>These results strongly indicate that soil arsenic contamination can be detected with Hyperion satellite hyperspectral data when combined with preprocessing and machine learning.</td></tr>
</table>

 doi:10.20944/preprints202503.0068.v1

<table>
<tr>
<td></td>
<td></td>
<td>(MLP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>
<table>
<tr><td>RFC</td><td>0.639</td><td>0.678</td><td>0.649</td><td>0.737</td><td>0.252</td></tr>
<tr><td>MLP</td><td>0.693</td><td>0.728</td><td>0.704</td><td>0.772</td><td>0.279</td></tr>
</table>
</td>
<td></td>
</tr>
<tr>
<td>[59]</td>
<td>Landsat 7 (ETM+)</td>
<td>Cubist (Cu), Random Forest (RF), Regression Tree (RT), Multiple Linear Regression (MLR).</td>
<td>Soil organic carbon (SOC), calcium carbonate equivalent (CCE), and clay content.</td>
<td>Total of 334 soil samples were collected from 0 to 30 cm depth.</td>
<td>334 soil samples</td>
<td>RMSE, R2 and RMSE%.</td>
<td>According to the RMSE and R2, Cu and RF resulted in the most accurate predictions for CCE and clay contents respectively, while both of RF and Cu models showed the highest performance to predict SOC content
<table>
<tr><th>Soil properties</th><th>Best Model</th><th>RMSE validation</th><th>R2 validation</th><th>RMSE calibration</th><th>R2 calibration</th></tr>
<tr><td>SOC</td><td>Cu and RF</td><td>0.34</td><td>0.55</td><td>0.14</td><td>0.93</td></tr>
<tr><td>CCE</td><td>RF</td><td>9.96</td><td>0.23</td><td>4.56</td><td>0.89</td></tr>
<tr><td>Clay</td><td>RF</td><td>7.86</td><td>0.15</td><td>3.53</td><td>0.92</td></tr>
</table>
</td>
<td>Results showed that remote sensing covariates (Ratio Vegetation Index and band 4) were the most important variables to explain the variability of SOC and CCE content, but only topographic attributes were responsible for clay content variation.</td>
</tr>
</table>

 doi:10.20944/preprints202503.0068.v1

<table>
<tr>
<td>[60]</td>
<td>Sentinel 2, Landsat 8</td>
<td>Neural Networks (NNs) and random forests (RF).</td>
<td>Crop productivity</td>
<td>GPP data directly from the principal Investigators of the sites, and integrated half-hourly data to daily GPP values, which were then used as the reference value for the validation of our GPP model.</td>
<td>-</td>
<td>R2, RMSE</td>
<td>GPP data. Our final neural network model is able to estimate GPP at the tested flux tower sites with r2 of 0.92 and RMSE of 1.38 g C d−1m−2, which outperforms empirical models based on vegetation indices.
<table>
<tr><th>Model</th><th>Properties</th><th>GPP r2 tes</th><th>LAI r2 test</th><th>GPP r2 val</th><th>GPP RMSE val.</th></tr>
<tr><td>NN (hidden layers)</td><td>(20,12)</td><td>0.92</td><td>0.62</td><td>0.88</td><td>1.38</td></tr>
<tr><td>NN (hidden layers)</td><td>(40,20,12)</td><td>0.95</td><td>0.68</td><td>0.91</td><td>1.41</td></tr>
<tr><td>RF (settings)</td><td>SW v.1</td><td>0.90</td><td>0.51</td><td>0.87</td><td>1.70</td></tr>
<tr><td>RF (settings)</td><td>SW v.2</td><td>0.90</td><td>0.51</td><td>0.89</td><td>1.58</td></tr>
</table>
SW v.1, increased the sample weight of data points with GPP below 2 µmol CO2 m−2 s
−1 to 2. SW v.2 d for the settings SW v.2 to 20.</td>
<td>The model successfully estimates gross primary productivity (GPP) across a variety of C3 crop types and environmental conditions even though it does not use any local information from the corresponding sites. This highlights its potential to map crop productivity from new satellite sensors at a global scale with the help of current Earth observation cloud computing platforms.</td>
</tr>
<tr>
<td>[61]</td>
<td>Sentinel 2, Sentinel 3</td>
<td>Data Mining Sharpener (DMS) based on an ensemble of decision-tree regressors.</td>
<td>Evaporation</td>
<td>land-cover map was based on Corine Land Cover (CLC) 2012 version 18.5, downloaded from the</td>
<td>-</td>
<td>RMSE, Bias, CV, r.</td>
<td>The correlation between the sensible heat fluxes is significantly lower (0.67) which leads to the correlation between EF estimates to lie between the two values (0.79). It is not immediately clear why the correlation of H is so much lower than that of the other fluxes (correlations of G and Rn are 0.99 and 0.90 respectively).</td>
<td>The results show that the fluxes derived with sharpened thermal data are of acceptable accuracy (relative error</td>
</tr>
</table>

 doi:10.20944/preprints202503.0068.v1

				Copernicus Land Monitoring Service and **meteorological data**, which in this study are obtained from the ERA-Interim reanalysis data set produced by the European Centre for Medium-Range Weather Forecasts (ECMWF).				lower than 20%) and provide more information at flux-tower footprint scale than the corresponding low-resolution fluxes.
[62]	Landsat 5, Landsat 8.	Random forest (RF) classifier.	Soil salinity	As ground truth we used the WoSIS Soil Profile Database, which is maintained by ISRIC – World Soil Information and includes over 100,000 georeferenced soil profiles. For the study the upper layer of soil profiles were selected for	In total, 15,188 data points were selected and used in further analysis	confusion matrix, overall accuracy, user's accuracy, producer's accuracy	The validation accuracy of the resulting maps was in the range of 67–70%.	It concludes that combining soil properties maps and thermal infrared imagery allows mapping of soil salinity development in space and time on a global scale

 doi:10.20944/preprints202503.0068.v1

<table>
<tr><td></td><td></td><td></td><td></td><td>which electrical conductivity (ECe) values are available. The thickness of this layer varied from 0 to 5 cm to 0–60 cm.</td><td></td><td></td><td></td><td></td></tr>
<tr><td>[63]</td><td>Landsat 8, Sentinel 2</td><td>ANN</td><td>Soil erosion</td><td>Sampling locations were carefully selected on the basis the most representative land cover, and overall conditions of the topsoil (about 0-20 depth).</td><td>30 surface soil samples</td><td>R2, RMSE, ordinary least square regression (OLSR) and geographical weighted regression (GWR)</td><td>The high corresponding R2 values (67%) for OLSR denoted the potential of field spectroscopy to describe soil health effectively.
<table>
<tr><th>Soil Erosion Parameter</th><th>Satellite type</th><th>Mean RMSE</th><th>Mean R2</th></tr>
<tr><td>SOM</td><td>Landsat 8</td><td>0.58</td><td>0.87</td></tr>
<tr><td>CaCO3</td><td>Landsat 8</td><td>8.08</td><td>0.79</td></tr>
<tr><td>K-factor</td><td>Landsat 8</td><td>0.0095</td><td>0.6</td></tr>
<tr><td>SOM</td><td>Sentinel-2</td><td>0.58</td><td>0.87</td></tr>
<tr><td>CaCO3</td><td>Sentinel-2</td><td>7.1</td><td>0.82</td></tr>
<tr><td>K-factor</td><td>Sentinel-2</td><td>0.0093</td><td>0.59</td></tr>
</table></td><td>The derived maps captured successfully the SOM, the CaCO3, and the K-factor spatial distribution in the GIS environment. The results may contribute to the design of erosion best management measures and wise land use planning in the study region. Finally, the results highlighted the fact that the terrain morphology is absolutely related to soil erosion rates rather than SOM values</td></tr>
</table>

 doi:10.20944/preprints202503.0068.v1

								that cannot successfully describe the soil erosion regime.
[64]	Landsat 8, Sentinel 2 MSI	Cubist model	Soil salinization	At each sampling point, four topsoil samples were collected and mixed (from 0 to 20 cm) using a soil drill. In the meantime, a portable GPS (UniStrong G120, positioning accuracy ≤5 m) was used to record the geographic locations.	64 topsoil samples in an arid desert region, the Ebinur Lake Wetland National Nature Reserve (ELWNNR)	R2, RMSE, NRMSE, RPD, RPIQ	The results showed that the measured soil salinity had a significant correlation with surface soil moisture (Pearson's r = 0.75). The introduction of TCW generated satisfactory estimating performance. Compared with OLI dataset, the combination of MSI dataset and Cubist model yielded overall better model performance and accuracy measures (R2=0.912, RMSE=6.462 dSm−1, NRMSE=9.226%, RPD=3.400 and RPIQ=6.824, respectively)	The differences between Landsat-8 OLI and Sentinel-2MSI were distinguishable. MSI image with finer spatial resolution performed better than OLI. Combining RS data sets and their derived TCW within a Cubist framework yielded accurate regional salinity map. The increased temporal revisiting frequency and spectral resolution of MSI data are expected to be positive enhancements to the acquisition of

 doi:10.20944/preprints202503.0068.v1

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>high-quality soil salinity information of desert soils.</td></tr>
<tr><td>[14]</td><td>Terra (Aster), Earth Observing-1 (EO-1) (Hyperion), Sentinel 2A, Landsat 8.</td><td>Artificial Neural Networks (ANN), Stepwise Multiple Linear Regression (SMLR) and PLSR.</td><td>Chromium (Cr)</td><td>The samples' mineralogy and Cr concentration were determined and were then subjected to laboratory reflectance spectroscopy in the range of Visible–Near Infrared–Shortwave Infrared (VNIR–SWIR: 350–2500 nm).</td><td>120 soil samples</td><td>R2, RMSE, RPD</td><td>Using satellite images, SD-SMLR provided the best prediction models with R2 values of 0.61 and 0.53 for Hyperion and Sentinel-2A, respectively.
<table>
<tr><th>Model</th><th>Satellite</th><th>R2</th><th>RMSE</th></tr>
<tr><td>SD-ANN</td><td>Aster</td><td>0.21</td><td>48.81</td></tr>
<tr><td>SD-ANN</td><td>Hyperion</td><td>0.61</td><td>25.82</td></tr>
<tr><td>SD-ANN</td><td>Sentinel-2A</td><td>0.33</td><td>36.29</td></tr>
<tr><td>SD-ANN</td><td>Landsat 8-OLI</td><td>0.24</td><td>45.06</td></tr>
<tr><td>SD-SMLR</td><td>Aster</td><td>0.31</td><td>42.63</td></tr>
<tr><td>SD-SMLR</td><td>Hyperion</td><td>0.68</td><td>23.61</td></tr>
<tr><td>SD-SMLR</td><td>Sentinel-2A</td><td>0.53</td><td>34.51</td></tr>
<tr><td>SD-SMLR</td><td>Landsat 8-OLI</td><td>0.45</td><td>40.58</td></tr>
<tr><td>SD-PLSR</td><td>Aster</td><td>0.22</td><td>45.37</td></tr>
<tr><td>SD-PLSR</td><td>Hyperion</td><td>0.54</td><td>29.11</td></tr>
<tr><td>SD-PLSR</td><td>Sentinel-2A</td><td>0.31</td><td>40.58</td></tr>
<tr><td>SD-PLSR</td><td>Landsat 8-OLI</td><td>0.24</td><td>41.55</td></tr>
</table></td><td>This study's findings indicated that applying the best prediction models obtained by spectroscopy to the selected wavebands of Hyperion and Sentinel-2A satellite imagery could be considered a promising technique for rapid, cost-effective and eco-friendly assessment of Cr concentration in highly heterogeneous mining areas.</td></tr>
</table>

 doi:10.20944/preprints202503.0068.v1

<table>
<tr>
<td>[65]</td>
<td>Sentinel 2</td>
<td>Support Vector Machine Regression (SVMR)</td>
<td>Soil Organic Carbon (SOC), Soil texture</td>
<td>The soil samples were taken at 0–10 cm depth as composite samples over an area of 6×6 m, air-dried, ground and sieved (≤2 mm) and thoroughly mixed before analyzing (ISO 11464:2006).</td>
<td>200 soil samples were collected using conditioned Latin Hypercube Sampling (cLHS) stratified random strategy</td>
<td>RMSEcv, RMSEp, RPD, Bias</td>
<td>The statistical accuracy attained using the LUCAS library was low and only the clay estimation model using Hyperion data showed suitable prediction accuracy with RMSE=7.98 and RPD=1.62. However, in PONMAC dataset, Sentinel- 2 simulated data provided the best results among the imagers for all properties except for silt.
<table>
<tr><th>Properties</th><th>RMSEcv</th><th>RMSEp</th><th>RPD</th><th>Bias</th></tr>
<tr><td>SOC</td><td>0.14</td><td>0.14</td><td>1.60</td><td>0.03</td></tr>
<tr><td>Clay</td><td>2.87</td><td>3.05</td><td>1.27</td><td>-0.44</td></tr>
<tr><td>Silt</td><td>5.71</td><td>6.28</td><td>1.13</td><td>-1.89</td></tr>
<tr><td>Sand</td><td>5.93</td><td>8.22</td><td>1.02</td><td>-0.95</td></tr>
</table>
</td>
<td>The SOC maps also confirmed that in areas with a high level of SOC, Sentinel-2 was able to detect SOC more precisely than the airborne sensors. However, a decrease in the model and map performances was clear in the case of parameters with low contents. The study also emphasized the importance of the super spectral Sentinel-2 data in soil characteristic assessments with a frequent revisit-time over larger areas than is currently done with laboratory and airborne instruments.</td>
</tr>
</table>

 doi:10.20944/preprints202503.0068.v1

[70]	Landsat 8 OLI	Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), Random Forest (RF), eXtreme Gradient Boosting (XGBoost)	Soil health indicators, including soil fauna, soil microbes, and soil organic matter (SOM)	Data collected from various sources including soil sample extractions, cultivation in controlled environments, and publicly available image databases like ImageNet	-	Accuracy, classification accuracy, R2 value	XGBoost model achieving nearly 100% accuracy for nematode strain Steinernema feltiae NY, with other strains above 97% accuracy. - Random Forest (RF) model trained on 168 soil samples achieved a linear relationship (r value) of 0.74 in SOM prediction using LIFS data. - PLS regression model demonstrating an R2 value of 0.84 for SOM prediction.	The study highlights the integration of ML and computer vision techniques with digital imaging and spectroscopy for soil health assessment. It demonstrates the potential of these technologies to enhance the accuracy and efficiency of soil health monitoring, emphasizing the need for comprehensive data sets and addressing challenges like environmental variability.

 doi:10.20944/preprints202503.0068.v1

[71]	Landsat 8/9	Stepwise Multiple Linear Regression (SMLR), Random Forest (RF)	Co, Cr, Cu, Fe, Mn, Ni, Pb, Zn	Samples collected from topsoil layers (0.3 m) of 19 soil pedons in the harrats arid region, Saudi Arabia	19 soil pedons	R2, RMSE, NRMSE	SMLR: Mean R2 varied between 0.38 (Zn) and 0.54 (Cu) with NRMSEs between 18.53% (Zn) and 26.03% (Cr). RF: Mean R2 ranged from 0.17 (Ni) to 0.40 (Cu) with NRMSEs between 19.15% (Co) and 27.76% (Mn)	The study demonstrated the capacity of SMLR to use environmental covariates (ECOVs) to predict heavy metals (HMs) concentrations and generate background levels. SMLR performed better than RF in predicting HMs. The established background levels are important for future environmental pollution and monitoring studies in the harrats arid region.
[79]	GaoFen-5 (GF-5)	Stacking model (PLSR, RFR, SVR)	Cd, As, Pb, Cu, Zn	Samples collected from topsoil layers (0-30 cm) using a grid pattern of 30 m by 30 m, totaling 415 samples	415 soil samples	R2, RMSE	Cd: R2 = 0.65, RMSE = 0.29; As: R2 = 0.60, RMSE = 5.19; Pb: R2 = 0.78, RMSE = 37.19; Cu: R2 = 0.85, RMSE = 4.76; Zn: R2 = 0.81, RMSE = 37.32	The study confirms that integrating geographical environmental factors (GEFs) into the SHMC prediction model significantly improves

 doi:10.20944/preprints202503.0068.v1

								prediction accuracy. The Stacking model demonstrated higher accuracy compared to single models, with notable performance improvements for Cd and As. The research emphasizes the potential of advanced hyperspectral remote sensing technology in environmental monitoring.
[83]	Pleiades	Deep Forest Algorithm	Tree counting	Satellite imagery from Pleiades for Kulon Progo district, Yogyakarta, Indonesia	-	F1 Score, Recall, Precision	Experiment 1: F1 = 0.760, Recall = 0.743, Precision = 0.778; Experiment 2: F1 = 0.774, Recall = 0.792, Precision = 0.756; Experiment 3: F1 = 0.779, Recall = 0.789, Precision = 0.769	The study successfully applied the Deep Forest algorithm to count trees using Pleiades satellite imagery. The best F1 score achieved was 0.779, indicating the algorithm's potential for accurate tree counting.

 doi:10.20944/preprints202503.0068.v1

[72]	Landsat 8 OLI, NOAA, ASTER-GDEM	Random Forest (RF)	Lead (Pb)	304 soil samples collected using a 2x2 km grid pattern, combined with multisource geographic data including historical and current satellite images	304 soil samples	R2, RMSE, RPIQ	R2 = 0.85, RMSE = 0.80 mg/kg, RPIQ = 4.09	The study developed a mapping method for soil potentially toxic elements (PTEs) using temporal-spatial-spectral (TSS) covariates combined with a random forest model. The model achieved high accuracy, demonstrating the importance of incorporating temporal parameters into soil PTE mapping for better environmental risk assessment and soil management.
[67]	Sentinel-1, Sentinel-2	• Random Forest (RF) • Support Vector Machine (SVM) • Multilayer Perceptron (MLP) • Convolutiona	• Land use and land cover (LULC) • Vegetation indices (NDVI, SAVI, NDBI, MNDWI, TCB)	Images from three different sensors, pre-processing, feature extraction, 10-fold cross-validation, test dataset	training and validation: 131 polygons, test data: 17 polygons	• Overall Accuracy (OA) • Kappa Index • Precision • Recall • Balanced Accuracy	•Convolutional Neural Network (CNN) with full dataset: • Overall Accuracy (OA): 0.969 • Kappa Index: 0.959 • Per Class Metrics (Precision, Recall, Balanced Accuracy): • Forest: Precision = 0.679, Recall = 1.000, Balanced Accuracy = 0.500 • Scrub: Precision = 1.000, Recall = 0.447, Balanced Accuracy = 0.724 • Dense Tree Crops (DTC): Precision = 0.988, Recall = 0.949, Balanced Accuracy = 0.881	The study demonstrated that the CNN model, when trained with a comprehensive dataset, provides superior accuracy for land use and

 doi:10.20944/preprints202503.0068.v1

		l Neural Network (CNN)					• Irrigated Grass Crops (IGC): Precision = 0.925, Recall = 0.996, Balanced Accuracy = 0.520 • Impermeable: Precision = 0.758, Recall = 0.920, Balanced Accuracy = 0.567 • Water: Precision = 1.000, Recall = 1.000, Balanced Accuracy = 0.835 • Bare Soil: Precision = 0.963, Recall = 0.852, Balanced Accuracy = 0.871 • Greenhouses: Precision = 0.996, Recall = 0.754, Balanced Accuracy = 0.420 • Netting: Precision = 0.286, Recall = 0.683, Balanced Accuracy = 0.314	land cover classification in semi-arid Mediterranean areas. However, the model's performance varies across different classes, highlighting the need for balanced training datasets to avoid overfitting.
[73]	Landsat-8 OLI	Random Forest (RF), Gradient Boosting Machine (GBM), Multi-layer Perceptron (MLP)	Soil salinity, Electrical conductivity (EC)	Samples collected using TDR-350 device for measuring EC, moisture, and temperature of soil, with 177 points collected around Maharloo Lake, 70% for training and 30% for testing	177	R^2, RMSE	GBM R^2 = 0.89, RMSE = 0.63; RF R^2 = 0.85, RMSE = 0.71; MLP R^2 = 0.75, RMSE = 0.88	The GBM model showed the best performance in predicting soil salinity, with the RF model also performing well, while the MLP model showed the worst performance. This model is highly effective for monitoring and managing soil salinity, particularly in arid and semi-arid regions.

 doi:10.20944/preprints202503.0068.v1

[68]	•Sentinel-2 •Planet Lab SuperDove (Synthetic SuperDove, SSD, and Actual SuperDove, ASD)	Random Forest Regression (RF), Support Vector Regression (SVR), Linear Regression, K-Nearest Neighbors (KNN), Decision Tree (DT)	Geological formations, Land cover types, Vegetation indices (NDVI, NDWI)	Ground truth data was collected using a combination of direct field observations and existing geological survey data. This involved the use of handheld GPS devices to mark the exact locations for sample collection, ensuring the accuracy of the spatial data used for training and validating the machine learning models. The process also included the use of various sensors to measure specific environmental parameters, which were then correlated with the	N/A	• R^2 (Coefficient of Determination) • RMSE (Root Mean Square Error) • MAE (Mean Absolute Error)	Best Performance Results: • Random Forest: • R^2: 0.92 • RMSE: 0.27 • MAE: 0.15 • Support Vector Regression (SVR): • R^2: 0.85 • RMSE: 0.32 • MAE: 0.19 • Decision Tree (DT): • R^2: 0.75 • RMSE: 0.45 • MAE: 0.28	The study demonstrated that the Random Forest algorithm, combined with Sentinel-2 and Planet Lab SuperDove imagery, provides highly accurate geological information extraction. This method proved superior to other machine learning models, offering significant potential for applications in geological mapping and environmental monitoring.

 doi:10.20944/preprints202503.0068.v1

				satellite imagery to enhance model training.				
[69]	Sentinel-2, PlanetScope (Dove satellites)	Random Forest Algorithm	Aboveground biomass, spectral reflectance	The study used a combination of destructive sampling, C-Dax pasture meters, and rising plate meters (RPM) to gather ground truth data across ten farms. The destructive sampling method involved physically harvesting the biomass from specific plots, while C-Dax pasture meters and RPM provided non-destructive measurements through	Estimated 12,000 field datasets	R^2, RMSE (Root Mean Squared Error), MAE (Mean Absolute Error)	Detailed Best Performance Results: • Sentinel-2: • R^2: 0.87 • RMSE: 439 kg DM/ha • MAE: 255 kg DM/ha • Synthetic SuperDove: • R^2: 0.92 • RMSE: 346 kg DM/ha • MAE: 208 kg DM/ha	The integration of Sentinel-2 and Planet SuperDove imagery with a Random Forest algorithm significantly improved the accuracy of pasture biomass estimation. This enhanced model supports more effective pasture management, especially in regions with frequent cloud cover, by providing timely and accurate biomass assessments.

 doi:10.20944/preprints202503.0068.v1

				reflectance and height estimations, respectively.				
[80]	PRISMA	Random Forest (RF), Support Vector Regression (SVR), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU)	Chlorophyll-a (Chl-a) concentrations	Data from three sub-alpine lakes, complemented by low-resolution Chlorophyll-a concentration maps	Data from three sub-alpine lakes, complemented by low-resolution Chlorophyll-a concentration maps	Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)	• SVR-4 (SVR): • Overall MAE: 0.687 µg/L • Overall RMSE: 0.895 µg/L • RF-10 (RF): • Overall MAE: 0.915 µg/L • Overall RMSE: 1.099 µg/L • GRU-8 (GRU): • Overall MAE: 1.186 µg/L • Overall RMSE: 1.321 µg/L • LSTM-13 (LSTM): • Overall MAE: 1.211 µg/L • Overall RMSE: 1.345 µg/L	The study demonstrated that the SVR model with standard scaling and PCA achieved the best performance in predicting Chlorophyll-a concentrations from PRISMA hyperspectral imagery. Enhancements in spatial resolution from Sentinel-3 to PRISMA were successfully achieved, though models tended to underestimate high Chl-a concentrations, suggesting the

 doi:10.20944/preprints202503.0068.v1

								need for additional PRISMA data acquisitions to improve accuracy.
[76]	PlanetScope (Dove satellites)	Support Vector Machines (SVMs), Decision Trees (DTs), Random Forests (RFs), Normal Bayes (NB), Artificial Neural Networks (ANNs)	Land use and land cover (LULC)	Empirical experiments, data representation, and pre-processing of satellite images	105 geo-referenced images	Precision, Recall, F-score, Kappa index	ANN classification accuracy: •ANN Precision: 0.9821 •ANN Recall: 0.9871 •ANN F-score: 0.9622 •ANN Kappa: 0.971	ANN achieved the highest accuracy for LULC classification, demonstrating the effectiveness of integrating multi-spectral satellite imagery with ML algorithms in Egypt.

 doi:10.20944/preprints202503.0068.v1

References

[1] N. J. Jiang et al., "Bio-mediated soil improvement: An introspection into processes, materials, characterization and applications," *Soil Use and Management*, vol. 38, no. 1. John Wiley and Sons Inc, pp. 68–93, Jan. 01, 2022. doi: 10.1111/sum.12736.

[2] E. Havugimana, B. Bhople, and A. Kumar, "Soil Pollution-Major Sources and Types of Soil Pollutants Integrated nutrient management View project Cluster frontline demonstration on pulses View project," 2017. [Online]. Available: https://www.researchgate.net/publication/321526846

[3] P. L. Lallas, "The Stockholm Convention on Persistent Organic Pollutants," *American Journal of International Law*, vol. 95, no. 3, pp. 692–708, Jul. 2001, doi: 10.2307/2668517.

[4] M. S. Askari, S. M. O'Rourke, and N. M. Holden, "Evaluation of soil quality for agricultural production using visible–near-infrared spectroscopy," *Geoderma*, vol. 243–244, pp. 80–91, Apr. 2015, doi: 10.1016/j.geoderma.2014.12.012.

[5] E. W. Slessarev et al., "Water balance creates a threshold in soil pH at the global scale," *Nature*, vol. 540, no. 7634, pp. 567–569, Dec. 2016, doi: 10.1038/nature20139.

[6] B. Hu et al., "Application of portable XRF and VNIR sensors for rapid assessment of soil heavy metal pollution," *PLoS One*, vol. 12, no. 2, p. e0172438, Feb. 2017, doi: 10.1371/journal.pone.0172438.

[7] A. Horta et al., "Potential of integrated field spectroscopy and spatial analysis for enhanced assessment of soil contamination: A prospective review," *Geoderma*, vol. 241–242, pp. 180–209, Mar. 2015, doi: 10.1016/j.geoderma.2014.11.024.

[8] L. Liu, W. Li, W. Song, and M. Guo, "Remediation techniques for heavy metal-contaminated soils: Principles and applicability," *Science of The Total Environment*, vol. 633, pp. 206–219, Aug. 2018, doi: 10.1016/j.scitotenv.2018.03.161.

[9] X. Jia, D. O'Connor, Z. Shi, and D. Hou, "VIRS based detection in combination with machine learning for mapping soil pollution," *Environmental Pollution*, vol. 268, p. 115845, Jan. 2021, doi: 10.1016/J.ENVPOL.2020.115845.

[10] T. Shi et al., "Proximal and remote sensing techniques for mapping of soil contamination with heavy metals," *Appl Spectrosc Rev*, vol. 53, no. 10, pp. 783–805, Nov. 2018, doi: 10.1080/05704928.2018.1442346.

[11] K. Tan et al., "Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning," *J Hazard Mater*, vol. 401, p. 123288, Jan. 2021, doi: 10.1016/J.JHAZMAT.2020.123288.

[12] S. Chakraborty et al., "Rapid assessment of regional soil arsenic pollution risk via diffuse reflectance spectroscopy," *Geoderma*, vol. 289, pp. 72–81, Mar. 2017, doi: 10.1016/j.geoderma.2016.11.024.

[13] X. Zhu, L. Cao, and Y. Liang, "Spatial distribution and risk assessment of heavy metals inside and outside a typical lead-zinc mine in southeastern China," *Environmental Science and Pollution Research*, vol. 26, no. 25, pp. 26265–26275, Sep. 2019, doi: 10.1007/s11356-019-05724-8.

[14] V. Khosravi, F. Doulati Ardejani, A. Gholizadeh, and M. Saberioon, "Satellite imagery for monitoring and mapping soil chromium pollution in a mine waste dump," *Remote Sens (Basel)*, vol. 13, no. 7, Apr. 2021, doi: 10.3390/rs13071277.

[15] L. Zhao et al., "Estimation Methods for Soil Mercury Content Using Hyperspectral Remote Sensing," *Sustainability*, vol. 10, no. 7, p. 2474, Jul. 2018, doi: 10.3390/su10072474.

[16] S. Zhang et al., "Hyperspectral inversion of heavy metal content in reclaimed soil from a mining wasteland based on different spectral transformation and modeling methods," *Spectrochim Acta A Mol Biomol Spectrosc*, vol. 211, pp. 393–400, Mar. 2019, doi: 10.1016/j.saa.2018.12.032.

[17] L. Wei, H. Pu, Z. Wang, Z. Yuan, X. Yan, and L. Cao, "Estimation of Soil Arsenic Content with Hyperspectral Remote Sensing," *Sensors*, vol. 20, no. 14, p. 4056, Jul. 2020, doi: 10.3390/s20144056.

[18] P. Liu et al., "Integrating a Hybrid Back Propagation Neural Network and Particle Swarm Optimization for Estimating Soil Heavy Metal Contents Using Hyperspectral Data," *Sustainability*, vol. 11, no. 2, p. 419, Jan. 2019, doi: 10.3390/su11020419.

[19] V. Khosravi, F. Doulati Ardejani, S. Yousefi, and A. Aryafar, "Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods," *Geoderma*, vol. 318, pp. 29–41, May 2018, doi: 10.1016/j.geoderma.2017.12.025.

[20] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community," *J Appl Remote Sens*, vol. 11, no. 04, p. 1, Sep. 2017, doi: 10.1117/1.jrs.11.042609.

[21] K. Fang, M. Pan, and C. Shen, "The Value of SMAP for Long-Term Soil Moisture Estimation With the Help of Deep Learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 2221–2233, Apr. 2019, doi: 10.1109/TGRS.2018.2872131.

[22] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Physica D*, vol. 404, p. 132306, Mar. 2020, doi: 10.1016/j.physd.2019.132306.

[23] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens Environ*, vol. 241, p. 111716, May 2020, doi: 10.1016/j.rse.2020.111716.

[24] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[25] K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.1078

[26] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," Aug. 2015, [Online]. Available: http://arxiv.org/abs/1508.01991

[27] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[28] A. C. Tricco et al., "A scoping review on the conduct and reporting of scoping reviews," *BMC Med Res Methodol*, vol. 16, no. 1, p. 15, Dec. 2016, doi: 10.1186/s12874-016-0116-4.

[29] A. C. Tricco et al., "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation," *Ann Intern Med*, vol. 169, no. 7, pp. 467–473, Oct. 2018, doi: 10.7326/M18-0850.

[30] J. P. T. Higgins et al., "The Cochrane Collaboration's tool for assessing risk of bias in randomised trials," *BMJ*, vol. 343, no. oct18 2, pp. d5928–d5928, Oct. 2011, doi: 10.1136/bmj.d5928.

[31] W. de S. Mendes et al., "A remote sensing framework to map potential toxic elements in agricultural soils in the humid tropics," *Environmental Pollution*, vol. 292, Jan. 2022, doi: 10.1016/j.envpol.2021.118397.

[32] H. Ma and S. Liang, "Development of the GLASS 250-m leaf area index product (version 6) from MODIS data using the bidirectional LSTM deep learning model," *Remote Sens Environ*, vol. 273, May 2022, doi: 10.1016/j.rse.2022.112985.

[33] X. Wang, C. Liu, and H. J. van der Fels-Klerx, "Regional prediction of multi-mycotoxin contamination of wheat in Europe using machine learning," *Food Research International*, vol. 159, Sep. 2022, doi: 10.1016/j.foodres.2022.111588.

[34] Y. Fang, L. Xu, A. Wong, and D. A. Clausi, "Multi-Temporal Landsat-8 Images for Retrieval and Broad Scale Mapping of Soil Copper Concentration Using Empirical Models," *Remote Sens (Basel)*, vol. 14, no. 10, May 2022, doi: 10.3390/rs14102311.

[35] F. Löw, K. Stieglitz, and O. Diemar, "Terrestrial oil spill mapping using satellite earth observation and machine learning: A case study in South Sudan," *J Environ Manage*, vol. 298, Nov. 2021, doi: 10.1016/j.jenvman.2021.113424.

[36] A. Sharma, X. Liu, and X. Yang, "Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks," *Neural Networks*, vol. 105. Elsevier Ltd., pp. 346–355, Sep. 01, 2018. doi: 10.1016/j.neunet.2018.05.019.

[37] É. de C. Paes, G. V. Veloso, A. A. da Fonseca, E. I. Fernandes-Filho, M. P. F. Fontes, and E. M. B. Soares, "Predictive modeling of contents of potentially toxic elements using morphometric data, proximal sensing, and chemical and physical properties of soils under mining influence," *Science of the Total Environment*, vol. 817, Apr. 2022, doi: 10.1016/j.scitotenv.2022.152972.

[38] D. Radočaj, M. Jurišić, M. Gašparović, I. Plaščak, and O. Antonić, "Cropland suitability assessment using satellite-based biophysical vegetation properties and machine learning," *Agronomy*, vol. 11, no. 8, Aug. 2021, doi: 10.3390/agronomy11081620.

[39] F. Pech-May, R. Aquino-Santos, G. Rios-Toledo, and J. P. F. Posadas-Durán, "Mapping of Land Cover with Optical Images, Supervised Algorithms, and Google Earth Engine," *Sensors*, vol. 22, no. 13, Jul. 2022, doi: 10.3390/s22134729.

 doi:10.20944/preprints202503.0068.v1

[40] L. Guo et al., "Mapping soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas," *Geoderma*, vol. 398, Sep. 2021, doi: 10.1016/j.geoderma.2021.115118.

[41] C. Xing, N. Chen, X. Zhang, and J. Gong, "A machine learning based reconstruction method for satellite remote sensing of soil moisture images with in situ observations," *Remote Sens (Basel)*, vol. 9, no. 5, May 2017, doi: 10.3390/rs9050484.

[42] X. Jia and D. Hou, "Mapping soil arsenic pollution at a brownfield site using satellite hyperspectral imagery and machine learning," *Science of the Total Environment*, vol. 857, Jan. 2023, doi: 10.1016/j.scitotenv.2022.159387.

[43] Y. Chen et al., "Microplastics pollution in the soil mulched by dust-proof nets: A case study in Beijing, China," *Environmental Pollution*, vol. 275, Apr. 2021, doi: 10.1016/j.envpol.2021.116600.

[44] S. Chen et al., "A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution," *Science of the Total Environment*, vol. 655, pp. 273–283, Mar. 2019, doi: 10.1016/j.scitotenv.2018.11.230.

[45] T. Shi et al., "Digital mapping of zinc in urban topsoil using multisource geospatial data and random forest," *Science of the Total Environment*, vol. 792, Oct. 2021, doi: 10.1016/j.scitotenv.2021.148455.

[46] N. Lin, R. Jiang, G. Li, Q. Yang, D. Li, and X. Yang, "Estimating the heavy metal contents in farmland soil from hyperspectral images based on Stacked AdaBoost ensemble learning," *Ecol Indic*, vol. 143, Oct. 2022, doi: 10.1016/j.ecolind.2022.109330.

[47] L. Wang, Y. Zhou, J. Liu, Y. Liu, Q. Zuo, and Q. Li, "Exploring the potential of multispectral satellite images for estimating the contents of cadmium and lead in cropland: The effect of the dimidiate pixel model and random forest," *J Clean Prod*, vol. 367, Sep. 2022, doi: 10.1016/j.jclepro.2022.132922.

[48] S. Zhou, H. Kaufmann, N. Bohn, M. Bochow, T. Kuester, and K. Segl, "Identifying distinct plastics in hyperspectral experimental lab-, aircraft-, and satellite data using machine/deep learning methods trained with synthetically mixed spectral data," *Remote Sens Environ*, vol. 281, Nov. 2022, doi: 10.1016/j.rse.2022.113263.

[49] R. Papi, S. Attarchi, A. Darvishi Boloorani, and N. Neysani Samany, "Knowledge discovery of Middle East dust sources using Apriori spatial data mining algorithm," *Ecol Inform*, vol. 72, Dec. 2022, doi: 10.1016/j.ecoinf.2022.101867.

[50] C. Acuña-Alonso, A. Novo, J. L. Rodríguez, S. Varandas, and X. Álvarez, "Modelling and evaluation of land use changes through satellite images in a multifunctional catchment: Social, economic and environmental implications," *Ecol Inform*, vol. 71, Nov. 2022, doi: 10.1016/j.ecoinf.2022.101777.

[51] K. Azizi, S. Ayoubi, K. Nabiollahi, Y. Garosi, and R. Gislum, "Predicting heavy metal contents by applying machine learning approaches and environmental covariates in west of Iran," *J Geochem Explor*, vol. 233, Feb. 2022, doi: 10.1016/j.gexplo.2021.106921.

[52] Y. Cheng et al., "Regional metal pollution risk assessment based on a long short-term memory model: A case study of the South Altai Mountain mining area, China," *J Clean Prod*, vol. 379, Dec. 2022, doi: 10.1016/j.jclepro.2022.134755.

[53] F. Derakhshan-Babaei, F. Mirchooli, M. Mohammadi, K. Nosrati, and M. Egli, "Tracking the origin of trace metals in a watershed by identifying fingerprints of soils, landscape and river sediments," *Science of the Total Environment*, vol. 835, Aug. 2022, doi: 10.1016/j.scitotenv.2022.155583.

[54] Z. Alvyar, F. Shahbazi, S. Oustan, O. Dengiz, and B. Minasny, "Digital mapping of potentially toxic elements enrichment in soils of Urmia Lake due to water level decline," *Science of the Total Environment*, vol. 808, Feb. 2022, doi: 10.1016/j.scitotenv.2021.152086.

[55] B. Zhang, B. Guo, B. Zou, W. Wei, Y. Lei, and T. Li, "Retrieving soil heavy metals concentrations based on GaoFen-5 hyperspectral satellite image at an opencast coal mine, Inner Mongolia, China," *Environmental Pollution*, vol. 300, May 2022, doi: 10.1016/j.envpol.2022.118981.

[56] J. P. D. Simioni, L. A. Guasselli, G. G. de Oliveira, L. F. C. Ruiz, and G. de Oliveira, "A comparison of data mining techniques and multi-sensor analysis for inland marshes delineation," *Wetl Ecol Manag*, vol. 28, no. 4, pp. 577–594, Aug. 2020, doi: 10.1007/s11273-020-09731-2.

[57] H. Teng, R. A. Viscarra Rossel, Z. Shi, T. Behrens, A. Chappell, and E. Bui, "Assimilating satellite imagery and visible-near infrared spectroscopy to model and map soil loss by water erosion in Australia," *Environmental Modelling and Software*, vol. 77, pp. 156–167, Mar. 2016, doi: 10.1016/j.envsoft.2015.11.024.

[58] A. Agrawal and M. R. Petersen, "Detecting arsenic contamination using satellite imagery and machine learning," *Toxics*, vol. 9, no. 12, Dec. 2021, doi: 10.3390/toxics9120333.

[59] M. Zeraatpisheh, S. Ayoubi, A. Jafari, S. Tajik, and P. Finke, "Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran," *Geoderma*, vol. 338, pp. 445–452, Mar. 2019, doi: 10.1016/j.geoderma.2018.09.006.

[60] A. Wolanin et al., "Estimating crop primary productivity with Sentinel-2 and Landsat 8 using machine learning methods trained with radiative transfer simulations," *Remote Sens Environ*, vol. 225, pp. 441–457, May 2019, doi: 10.1016/j.rse.2019.03.002.

[61] R. Guzinski and H. Nieto, "Evaluating the feasibility of using Sentinel-2 and Sentinel-3 satellites for high-resolution evapotranspiration estimations," *Remote Sens Environ*, vol. 221, pp. 157–172, Feb. 2019, doi: 10.1016/j.rse.2018.11.019.

[62] K. Ivushkin, H. Bartholomeus, A. K. Bregt, A. Pulatov, B. Kempen, and L. de Sousa, "Global mapping of soil salinity change," *Remote Sens Environ*, vol. 231, Sep. 2019, doi: 10.1016/j.rse.2019.111260.

[63] D. D. Alexakis, E. Tapoglou, A. E. K. Vozinaki, and I. K. Tsanis, "Integrated use of satellite remote sensing, artificial neural networks, field spectroscopy, and GIS in estimating crucial soil parameters in terms of soil erosion," *Remote Sens (Basel)*, vol. 11, no. 9, May 2019, doi: 10.3390/rs11091106.

[64] J. Wang et al., "Machine learning-based detection of soil salinity in an arid desert region, Northwest China: A comparison between Landsat-8 OLI and Sentinel-2 MSI," *Science of the Total Environment*, vol. 707, Mar. 2020, doi: 10.1016/j.scitotenv.2019.136092.

[65] A. Gholizadeh, D. Žižala, M. Saberioon, and L. Borůvka, "Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging," *Remote Sens Environ*, vol. 218, pp. 89–103, Dec. 2018, doi: 10.1016/j.rse.2018.09.015.

[66] M. J. Page et al., "Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement," *J Clin Epidemiol*, vol. 134, pp. 103–112, Jun. 2021, doi: 10.1016/j.jclinepi.2021.02.003.

[67] F. Alonso-Sarría, C. Valdivieso-Ros, and F. Gomariz-Castillo, "Analysis of the hyperparameter optimisation of four machine learning satellite imagery classification methods," *Comput Geosci*, 2024, doi: 10.1007/s10596-024-10285-y.

[68] A. Yadav, S. Saraswat, and N. Faujdar, "Geological Information Extraction from Satellite Imagery Using Machine Learning," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICRITO56286.2022.9964623.

[69] M. G. Ogungbuyi et al., "Improvement of pasture biomass modelling using high-resolution satellite imagery and machine learning," *J Environ Manage*, vol. 356, Apr. 2024, doi: 10.1016/j.jenvman.2024.120564.

[70] L. Wang et al., "Advancing Soil Health: Challenges and Opportunities in Integrating Digital Imaging, Spectroscopy, and Machine Learning for Bioindicator Analysis," *Analytical Chemistry*. American Chemical Society, May 21, 2023. doi: 10.1021/acs.analchem.3c05311.

[71] M. M. Sulieman, F. Kaya, A. Keshavarzi, A. M. Hussein, A. S. Al-Farraj, and E. C. Brevik, "Spatial variability of some heavy metals in arid harrats soils: Combining machine learning algorithms and synthetic indexes based-multitemporal Landsat 8/9 to establish background levels," *Catena (Amst)*, vol. 234, Jan. 2024, doi: 10.1016/j.catena.2023.107579.

[72] X. Xu, Z. Wang, X. Song, W. Zhan, and S. Yang, "A remote sensing-based strategy for mapping potentially toxic elements of soils: Temporal-spatial-spectral covariates combined with random forest," *Environ Res*, vol. 240, Jan. 2024, doi: 10.1016/j.envres.2023.117570.

[73] S. Alamdar, F. Ghazban, and A. Zarei, "EFFICIENCY OF MACHINE LEARNING ALGORITHMS IN SOIL SALINITY DETECTION USING LANDSAT-8 OLI IMAGERY," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus Publications, Jan. 2023, pp. 49–55. doi: 10.5194/isprs-annals-X-4-W1-2022-49-2023.

 doi:10.20944/preprints202503.0068.v1

[74] A. Gholizadeh, D. Žižala, M. Saberioon, and L. Borůvka, "Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging," *Remote Sens Environ*, vol. 218, pp. 89–103, Dec. 2018, doi: 10.1016/j.rse.2018.09.015.

[75] K. Azizi, S. Ayoubi, K. Nabiollahi, Y. Garosi, and R. Gislum, "Predicting heavy metal contents by applying machine learning approaches and environmental covariates in west of Iran," *J Geochem Explor*, vol. 233, p. 106921, Feb. 2022, doi: 10.1016/j.gexplo.2021.106921.

[76] R. Mahmoud, M. Hassanin, H. Al Feel, and R. M. Badry, "Machine Learning-Based Land Use and Land Cover Mapping Using Multi-Spectral Satellite Imagery: A Case Study in Egypt," *Sustainability (Switzerland)*, vol. 15, no. 12, Jun. 2023, doi: 10.3390/su15129467.

[77] https://sentinel.esa.int/web/sentinel/missions/sentinel-2, "European Space Agency. 'Sentinel-2.' Sentinel Online."

[78] "USGS. 'Landsat.' Core Science Systems, National Land Imaging Program, United States Geological Survey. Accessed May 11, 2023. ," https://www.usgs.gov/core-science-systems/nli/landsat.

[79] L. Yao, M. Xu, Y. Liu, R. Niu, X. Wu, and Y. Song, "Estimating of heavy metal concentration in agricultural soils from hyperspectral satellite sensor imagery: Considering the sources and migration pathways of pollutants," *Ecol Indic*, vol. 158, Jan. 2024, doi: 10.1016/j.ecolind.2023.111416.

[80] J. F. Amieva, D. Oxoli, and M. A. Brovelli, "Machine and Deep Learning Regression of Chlorophyll-a Concentrations in Lakes Using PRISMA Satellite Hyperspectral Imagery," *Remote Sens (Basel)*, vol. 15, no. 22, Nov. 2023, doi: 10.3390/rs15225385.

[81] X. Jia and D. Hou, "Mapping soil arsenic pollution at a brownfield site using satellite hyperspectral imagery and machine learning," *Science of the Total Environment*, vol. 857, Jan. 2023, doi: 10.1016/j.scitotenv.2022.159387.

[82] P. B. Tchounwou, C. G. Yedjou, A. K. Patlolla, and D. J. Sutton, "Heavy Metal Toxicity and the Environment," 2012, pp. 133–164. doi: 10.1007/978-3-7643-8340-4_6.

[83] S. Austen et al., "Tree Counting with Deep Forest Algorithm for Kulon Progo District in Yogyakarta, Indonesia Using Pleiades Satellite Imagery," 2024. [Online]. Available: https://www.researchgate.net/publication/378008284