



# Comprehensive Survey on Ensemble Machine Learning for Soil Health Monitoring

Ten Alexander Iosifovich<sup>1</sup>; Seungjae Lee<sup>2\*</sup>

<sup>1,2</sup> Computer Engineering Department, Sunmoon University, Korea

<sup>1</sup> [keshtenscript27@sunmoon.ac.kr](mailto:keshtenscript27@sunmoon.ac.kr); <sup>2</sup> [leeko@sunmoon.ac.kr](mailto:leeko@sunmoon.ac.kr)

**DOI:** <https://doi.org/10.47760/ijcsmc.2025.v14i12.002>

**Abstract:** Recent years have seen rapid progress in soil monitoring technologies as the demand for real-time, scalable, and data-driven assessment of soil health continues to rise. Advances across sensing systems, remote observation tools, digital soil infrastructures, and machine learning have reshaped the way soil physical, chemical, and biological properties are measured and interpreted. Studies have explored the limitations of conventional point-based measurements, highlighted the growing influence of environmental pressures on soil degradation, and emphasized the need for harmonized monitoring frameworks at regional and national scales. Emerging approaches—including IoT-enabled soil sensing, Earth observation for large-scale mapping, microbiome-based indicators, and machine learning-driven predictive modeling—offer new opportunities but also introduce technical, logistical, and data-related challenges. Given the rapid evolution and fragmentation of these developments, a comprehensive survey is necessary to consolidate current knowledge, compare methodological trends, and outline research directions that support reliable, continuous, and integrated soil monitoring systems for sustainable land management.

**Keywords:** Soil Monitoring, AI for agriculture, Soil Health, Machine Learning, IoT-Based Monitoring

## I. INTRODUCTION

Monitoring soil health has become an urgent global priority as agricultural intensification, climate variability, and pollution increasingly disrupt soil systems. Traditional soil assessment remains dominated by laboratory analysis and periodic sampling, yet these approaches are insufficient for capturing the spatial and temporal variability required for modern land management [1]. At the same time, long-term studies continue to show that anthropogenic pressures and climatic factors accelerate soil degradation, reinforcing the need for more systematic monitoring strategies [2]. Efforts to harmonize national and continental soil databases reveal

substantial inconsistencies between existing monitoring programs, demonstrating that variations in data collection and indicator definitions hinder cross-regional comparability and coordinated management [3]. Complementary work on soil–plant system modeling highlights that advancing soil health assessment requires integrating observational data with process-based understanding to better characterize physical, chemical, and biological dynamics [4]. Recent developments in monitoring infrastructures further illustrate the importance of resilient and adaptive systems. For example, in regions affected by conflict, disruptions to laboratory capacity, information systems, and regulatory frameworks have revealed the vulnerability of soil monitoring networks and the critical need for robust institutional support [5]. Technological innovation has simultaneously reshaped the monitoring landscape. IoT-based sensing platforms enable continuous, real-time soil measurement while addressing challenges around system authentication and data reliability [6]. Remote sensing has expanded the ability to characterize broad-scale soil conditions, including salinity, moisture, and degradation patterns, though persistent challenges remain regarding resolution, retrieval accuracy, and data processing complexity [7]. Additional research emphasizes the importance of monitoring environmentally stressed areas such as mining zones, where soil contamination and heavy metal accumulation require advanced diagnostic and remediation techniques [8]. Machine learning has emerged as a central component in modern soil monitoring pipelines. Studies integrating ML with IoT systems demonstrate improved nutrient assessment and decision support capabilities for precision agriculture [9], while other work combines ML with remote sensing to estimate soil indicators at regional scales and support environmental monitoring [10, 11]. ML has also advanced soil moisture estimation using satellite data, despite significant challenges related to meteorological variability, vegetation cover, and model generalization [12]. Beyond physical and chemical properties, microbiome-based approaches powered by supervised ML offer promising new biological indicators for soil health prediction [13]. Machine-learning-driven forecasting has also proven effective for modeling engineered soil behavior, including soil settlement under infrastructure development [15], illustrating the growing breadth of ML applications in soil analysis. Although these developments represent significant progress, current research remains fragmented across sensors, modeling frameworks, remote sensing technologies, and data-driven analytics. The rapid expansion of digital soil infrastructures, Earth observation tools, and ML-enabled inference has created a need for a unified perspective that synthesizes these advances and identifies persistent gaps. Therefore, a comprehensive survey is necessary to integrate the latest developments, evaluate their strengths and limitations, and outline future research directions for building robust, continuous, and scalable soil monitoring systems capable of supporting sustainable environmental and agricultural management.

## II. BACKGROUND AND MOTIVATION

### II.I. Soil monitoring

Soil monitoring has evolved into a critical scientific and environmental priority as soils face increasing pressures from climate change, land degradation, intensive agriculture, and industrial pollution. Traditional soil assessment methods—typically based on manual sampling and laboratory analysis—provide high accuracy but lack the temporal resolution required to capture rapid changes in soil conditions. This limitation has driven a shift toward real-time and continuous soil monitoring frameworks, which integrate sensors, remote sensing platforms, and modeling tools to provide dynamic and spatially extensive information on soil health parameters. Recent reviews highlight that continuous monitoring enables earlier detection of stressors, improved decision-making, and more efficient resource management; although challenges related to sensor calibration, data reliability, and system scalability remain significant [1].

Long-term monitoring efforts have also expanded in scope as researchers increasingly recognize the impact of anthropogenic activities and climatic variability on soil quality. Industrial emissions, agricultural intensification, land-use change, and extreme weather patterns have introduced complex temporal trends in soil degradation, requiring more robust and coordinated monitoring infrastructures [2]. In response, regional and continental initiatives—such as the LUCAS Soil program—have been working toward harmonized and standardized monitoring networks capable of integrating national datasets into unified frameworks, thereby improving data comparability across geopolitical regions [3].

Recent scientific advances further emphasize that soil health cannot be evaluated solely through physicochemical parameters. Instead, modern frameworks incorporate the soil–plant system as a coupled, dynamic environment in which biological, hydrological, and chemical processes interact. Reviews show that measuring soil moisture fluxes, nutrient cycling, plant responses, and subsurface processes contributes to a more holistic understanding of soil functioning and long-term resilience [4]. In some regions, geopolitical instability adds another layer of complexity: disruptions to monitoring infrastructures during conflict have exposed vulnerabilities in national soil databases and highlighted the need for resilient, distributed sensing systems [5].

Parallel to these developments, technological innovations have accelerated the use of IoT-based sensor networks, which enable automated, user-authenticated, and remotely accessible monitoring systems capable of tracking moisture, temperature, salinity, and nutrient levels in real time [6]. Remote sensing applications have

likewise expanded, especially for large-scale assessments such as salinization mapping, land degradation detection, and soil moisture estimation using multispectral and hyperspectral platforms [7,12]. These methods are now routinely combined with machine learning models, which enhance prediction accuracy and allow the estimation of unobservable indicators such as heavy-metal contamination, soil nutrient profiles, soil microbiome characteristics, and soil health indices [8–11,13–14]. Machine-learning approaches have even proven valuable in specialized engineering environments, where predictive models are used to forecast soil settlement and stability under construction-induced stress [15].

Together, these developments demonstrate that soil monitoring has transitioned from a static, laboratory-centered practice to a multi-modal, data-driven, and interdisciplinary field. This shift provides rich opportunities—and also significant complexity—which underscores the need for comprehensive surveys synthesizing advances in sensing technologies, remote-sensing platforms, machine-learning methods, and ensemble learning strategies.

## **II.II. Machine learning in soil monitoring**

Machine learning has rapidly become a central component of modern soil monitoring, complementing traditional sensing and laboratory methods by enabling faster, scalable, and more cost-effective analysis of complex soil datasets. Early ML applications in soil science focused on predicting individual parameters—such as moisture or nutrient concentrations—but recent advances have expanded the field toward multi-modal prediction frameworks, remote-sensing integration, microbiome-based soil health analytics, and hybrid IoT–ML systems. A major driver behind ML adoption is the increasing availability of soil measurements from continuous monitoring systems and remote sensing platforms, which generate large and heterogeneous datasets that exceed the analytical capacity of conventional statistical methods. Reviews show that integrating machine learning with continuous monitoring allows models to capture temporal variability in soil moisture, salinity, and physicochemical parameters more effectively than static sampling approaches [1,7,12]. ML-based systems have demonstrated improved ability to process sensor drift, noisy inputs, and environmental fluctuations, making them well-suited for real-time applications. One of the most active areas is soil nutrient estimation and crop recommendation, where ML models such as random forests, gradient boosting machines, and neural networks are used to infer macro- and micronutrient levels from IoT sensor data, environmental variables, and agricultural management history. These approaches help predict nutrient deficiencies and optimize fertilization strategies, while simultaneously reducing the need for costly laboratory analysis [9]. Machine learning also enhances remote-sensing workflows: the estimation of soil indicators—including organic carbon, moisture content, heavy-metal contamination, and salinization severity—now routinely incorporates models such as support vector machines, regression trees, k-nearest neighbors, deep learning architectures, and ensemble learners to transform hyperspectral, multispectral, and thermal imagery into soil condition maps [7,10,11]. In addition, biological data have opened new frontiers for ML. The soil microbiome, traditionally difficult to interpret due to its immense complexity, can now be leveraged to predict soil health metrics through supervised learning techniques. Studies demonstrate that machine learning can extract meaningful patterns from high-dimensional sequencing data to estimate biological activity, nutrient cycling potential, and overall soil vitality, providing insights that were previously inaccessible with conventional methods [13,14]. More specialized engineering and environmental applications also illustrate the versatility of ML. For example, predictive models have been used to forecast soil settlement during underground construction activities, where nonlinear subsurface interactions make mechanistic modeling challenging. ML-based approaches overcome these limitations by learning deformation patterns directly from historical monitoring data [15]. Despite these advances, the field faces notable challenges. Many soil datasets remain sparse, noisy, or imbalanced—particularly in large-scale monitoring systems—requiring more robust models capable of generalizing across diverse soil types, climate regimes, and land-use conditions. Furthermore, the integration of multi-source data (sensor networks, drones, satellites, laboratory analyses, and biological assays) still lacks standardized frameworks, limiting reproducibility and cross-regional comparability. As recent reviews emphasize, improving model interpretability, harmonizing datasets, and enhancing cross-domain fusion between physical, chemical, and biological indicators remain key priorities for the next generation of ML-driven soil monitoring systems [1–4,7,10,12]. Together, these trends indicate that machine learning has shifted from an auxiliary tool to a core analytical engine for soil monitoring—enabling dynamic prediction, scalable mapping, and biological insight at unprecedented resolution. The growing use of ensemble learning, hybrid sensing frameworks, and multi-modal feature fusion further underscores the need for comprehensive surveys to synthesize the rapidly expanding methodological landscape.\*\*\*\*\*

## **II.III. The Concept and potential of Ensemble machine learning in soil management**

Ensemble machine learning has emerged as one of the most powerful paradigms in modern data analytics, offering significant potential for advancing soil monitoring across diverse environmental, agricultural, and engineering applications. Unlike single-model approaches that rely on one algorithm to learn the underlying

structure of data, ensemble learning combines multiple models to improve robustness, stability, and predictive accuracy. This is especially important in soil science, where datasets are often noisy, limited in size, highly imbalanced, and characterized by non-linear relationships influenced by climate, land use, vegetation, and anthropogenic activity. Recent reviews of ensemble methods emphasize their superior ability to reduce variance, mitigate bias, and capture complex feature interactions, making them particularly suitable for the heterogeneous data environments typical of soil monitoring systems [16]. One of the defining strengths of ensemble learning is its adaptability to multi-source and multi-scale soil information. Modern soil monitoring increasingly relies on heterogeneous data streams such as laboratory assays, in situ sensors, remote sensing imagery, hydrological observations, and even microbiome sequencing. Each data type represents different spatial resolutions, temporal scales, and noise characteristics, which makes modeling soil properties with a single algorithm challenging. Ensemble techniques such as random forests, gradient boosting, bagging, stacking, and hybrid deep ensemble architectures are uniquely capable of integrating these diverse datasets. They provide a flexible framework for fusing weak learners or complementary models, enabling more comprehensive interpretation of soil moisture dynamics [12], salinity variation [7], heavy metal contamination [8], organic matter distribution [10], and biological activity indicated by microbiome structure [13,14]. Another advantage of ensemble learning in soil monitoring is its ability to enhance the reliability of real-time and continuous measurement systems. Continuous soil monitoring technologies, as reviewed in recent studies, produce temporal data streams that may include gaps, drifting sensors, inconsistent sampling frequencies, and abrupt environmental changes [1]. Ensemble models can manage these challenges by leveraging multiple predictive functions, each learning different temporal patterns or error structures. As a result, ensembles are more resilient to sudden fluctuations and measurement inconsistencies than individual algorithms. This stability is crucial for applications such as nutrient management and irrigation scheduling, where predictive accuracy directly influences crop productivity and resource efficiency [9]. The potential of ensemble learning becomes even more apparent in remote sensing-driven soil monitoring. Remote sensing data often include spectral signatures from satellites, drones, and airborne sensors, which require sophisticated modeling to extract meaningful soil information. Ensemble methods can combine the strengths of different learners to improve feature selection, reduce overfitting on high-dimensional spectral datasets, and enhance classification or regression performance when assessing soil indicators at large geographical scales [7,10,11]. Their capacity to integrate spatial context, environmental variables, and spectral indices makes them effective for mapping soil salinity, moisture, organic carbon, and contamination in a more consistent and interpretable manner. Ensemble machine learning also shows substantial promise for predictive modeling in soil dynamics and geotechnical processes. For example, forecasting soil settlement during underground construction requires capturing complex nonlinear behavior influenced by excavation rate, soil type, moisture, and stress redistribution. Ensemble approaches can combine neural networks, regression models, and decision trees to deliver more accurate forecasts and reduce prediction uncertainty, showing clear advantages over traditional single-model techniques [15]. Perhaps the most forward-looking potential lies in the integration of ensemble learning within emerging soil-plant digital twin frameworks. These systems combine physics-based models, machine learning, remote sensing, and data assimilation to simulate soil processes and crop responses in real time, providing a powerful tool for understanding soil health under changing environmental conditions [4]. Ensemble methods can enhance these digital twins by blending multiple mechanistic and data-driven models, thereby improving the reliability of predictions, especially when input data are sparse or uncertain. Overall, ensemble machine learning offers a compelling pathway to overcoming long-standing challenges in soil monitoring, including data scarcity, imbalance, high noise, spatial heterogeneity, and the need for multi-modal integration. As soil monitoring continues to evolve toward continuous, real-time, and digital-twin-enabled frameworks, ensembles provide the methodological foundation for building robust, adaptive, and high-performance analytical systems capable of supporting sustainable land management, environmental protection, and precision agriculture.

### III. POTENTIAL CHALLENGES OF ENSEMBLING

Despite the considerable advantages offered by ensemble machine learning, its application to soil monitoring is not without significant challenges and limitations that must be acknowledged for realistic and effective deployment. One of the most fundamental issues is the substantial computational cost associated with training and maintaining multiple models simultaneously. Ensemble approaches such as boosting, bagging, and stacking often require repeated training cycles, optimization of numerous hyperparameters, and storage of multiple learners, which can strain computational resources, especially when processing high-dimensional remote sensing imagery or large-scale continuous monitoring datasets. These computational demands become even more restrictive when ensembles are integrated with deep learning architectures, which already require intensive GPUs or cloud infrastructures that may be inaccessible in many agricultural or environmental monitoring settings. Data-related constraints also present major hurdles. Although ensemble learning can improve prediction accuracy in data-scarce environments, it still fundamentally depends on the availability of sufficiently

representative and diverse datasets to avoid overfitting or biased estimations. Soil data are often region-specific, temporally inconsistent, or collected using heterogeneous methods, which complicates ensemble training and risks producing models that generalize poorly across different soil types, climatic regions, or land-use systems. Even advanced ensemble methods may reinforce underlying data biases rather than correct them, especially when training datasets are imbalanced or lack samples from stressed or degraded soils. Moreover, many soil properties exhibit strong spatial heterogeneity, and ensembles trained without explicit spatial considerations may misinterpret local variability or amplify variance errors. Another challenge is the reduced interpretability associated with many ensemble models. While individual decision trees or simple regressors offer transparent decision pathways, ensembles combine numerous learners into a single predictive output, obscuring the reasoning behind predictions. This opacity can be problematic in soil monitoring applications where domain experts, policymakers, and land managers require clear explanations to justify interventions or regulatory decisions. For instance, decisions involving soil contamination, salinization risk, or nutrient deficiency assessments often necessitate transparency that complex ensembles struggle to provide. Although model interpretability tools exist, such as feature importance scores or SHAP values, they add analytical complexity and are not always intuitive for non-technical stakeholders. Integration challenges also arise when ensemble methods are applied to multi-source soil datasets. Combining laboratory measurements, in situ sensor streams, satellite imagery, environmental metadata, and biological indicators requires careful preprocessing, synchronization, and normalization. Misalignment in temporal resolution, spatial scale, or sampling methodology can cause inconsistencies that propagate through ensemble models, reducing their reliability. Data fusion from sources such as microbiome sequencing and remote sensing further complicates ensemble integration due to their vastly different dimensionalities and noise structures. A related limitation is the risk of diminishing returns. While ensembles typically outperform individual models, the improvement may be marginal in some soil monitoring tasks, especially when the base learners are highly correlated or when the underlying soil property is strongly influenced by external environmental factors that are not captured in the dataset. In such cases, increased model complexity may not justify the minimal performance gains, particularly for resource-limited agricultural environments. Model maintenance poses additional difficulties. Ensemble models deployed in real-time monitoring systems must continuously adapt to changing soil and climatic conditions, sensor drift, land-use modifications, and anthropogenic disturbances. Updating or retraining ensembles can be cumbersome, as it may require recalibrating multiple base learners or redesigning the fusion strategy, especially in systems that generate continuous data streams such as real-time moisture and nutrient monitoring [1]. These maintenance challenges become even more pronounced in digital twin frameworks, where ensembles must integrate seamlessly with physical models, data assimilation techniques, and real-time sensor networks [4]. Finally, operational challenges limit ensemble adoption in applied soil monitoring. Many agricultural practitioners, environmental agencies, and local monitoring groups lack the expertise required to implement and interpret sophisticated ML systems. Ensemble learning workflows involve complex preprocessing, feature engineering, hyperparameter tuning, and model validation procedures that may not align with the technical capacity of field-level users. Without accessible tools or standardized frameworks, the benefits of ensemble learning may remain confined to research settings rather than achieving practical uptake in soil management. In summary, while ensemble machine learning offers powerful advantages for soil monitoring, its challenges—including computational cost, data limitations, interpretability concerns, model maintenance burdens, and practical deployment difficulties—highlight the need for careful methodological design, transparent reporting, and application-specific optimization.

#### IV. CONCLUSION

Soil monitoring has entered a new era in which the integration of sensing technologies, remote observation, and advanced computational approaches is fundamentally reshaping our ability to understand, evaluate, and manage soil health. Across environmental science, agriculture, land restoration, and geotechnical engineering, rapid advancements in machine learning and sensor-based analytics are enabling more accurate, scalable, and real-time assessments of soil conditions than ever before. Within this evolving landscape, ensemble machine learning stands out as a particularly powerful methodology capable of addressing many of the long-standing challenges that have limited traditional soil monitoring approaches. By combining the strengths of multiple models, ensemble methods deliver enhanced predictive performance, improved robustness against noisy or incomplete data, and superior capacity to capture the nonlinear dynamics inherent in soil processes. This survey highlights how ensemble methods offer meaningful improvements across a wide spectrum of soil monitoring tasks, including moisture prediction, salinity assessment, nutrient evaluation, heavy metal contamination mapping, soil settlement forecasting, and biological health prediction. Their ability to integrate multi-source datasets—ranging from IoT sensors and laboratory measurements to satellite imagery and microbiome sequencing—positions them as a unifying analytical framework in a field increasingly defined by data heterogeneity and complexity. Ensemble learning not only strengthens the accuracy of soil property estimation but also contributes to more adaptive and resilient monitoring systems capable of operating under varying

environmental or operational conditions. At the same time, the review underscores that ensemble machine learning is not an automatic solution for every soil monitoring challenge. Computational demands, interpretability concerns, difficulties in multi-source data fusion, and the need for continuous model maintenance introduce practical barriers to widespread adoption. These limitations highlight the importance of thoughtful algorithm design, transparent workflows, and domain-specific calibration to ensure that ensemble-based approaches remain both effective and operationally feasible. As the field continues to evolve, future work should aim to streamline ensemble workflows, develop interpretability techniques that make model outputs accessible to non-experts, and design ensemble architectures tailored specifically to soil processes and environmental variability. Overall, the convergence of sensing technologies, remote observation systems, and ensemble machine learning offers an unprecedented opportunity to modernize soil monitoring practices. By harnessing diverse data streams and leveraging the complementary strengths of multiple algorithms, ensemble methods provide a pathway toward more accurate, intelligent, and resilient soil assessment frameworks. As global pressures on land resources intensify due to climate change, population growth, and industrial development, such advanced analytical capabilities will become indispensable for sustainable soil management and long-term environmental stewardship.

## REFERENCES

- [1]. Fan, Yingzheng, et al. "A critical review for real-time continuous soil monitoring: Advantages, challenges, and perspectives." *Environmental Science & Technology* 56.19 (2022).
- [2]. Bungau, Simona, et al. "Expatriating the impact of anthropogenic aspects and climatic factors on long-term soil monitoring and management." *Environmental Science and Pollution Research* 28.24 (2021).
- [3]. Froger, Claire, et al. "Comparing LUCAS Soil and national systems: Towards a harmonized European Soil monitoring network." *Geoderma* 449 (2024).
- [4]. Zeng, Yijian, et al. "Monitoring and modeling the soil-plant system toward understanding soil health." *Reviews of Geophysics* 63.1 (2025).
- [5]. Romanova, Svitlana, Yuriy Dmytruk, and Yaroslava Zhukova. "Soil monitoring infrastructure in response to war." *International Journal of Environmental Studies* 81.1 (2024):.
- [6]. Shukla, Varun, Shivani Dixit, and Prem Dixit. "An IoT Based User Authenticated Soil Monitoring System." *Adhoc & sensor wireless networks* 53 (2022).
- [7]. Sahbeni, Ghada, et al. "Challenges and opportunities in remote sensing for soil salinization mapping and monitoring: A review." *Remote Sensing* 15.10 (2023).
- [8]. Haghighizadeh, Atoosa, et al. "Comprehensive analysis of heavy metal soil contamination in mining Environments: Impacts, monitoring Techniques, and remediation strategies." *Arabian Journal of Chemistry* 17.6 (2024).
- [9]. Islam, Md Reazul, et al. "Machine learning enabled IoT system for soil nutrients monitoring and crop recommendation." *Journal of Agriculture and Food Research* 14 (2023).
- [10]. Diaz-Gonzalez, Freddy A., et al. "Machine learning and remote sensing techniques applied to estimate soil indicators—review." *Ecological Indicators* 135 (2022).
- [11]. Kim, Ye Inn, et al. "Applications of machine learning and remote sensing in soil and water conservation." *Hydrology* 11.11 (2024).
- [12]. Lamichhane, Manoj, Sushant Mehan, and Kyle R. Mankin. "Soil moisture prediction using remote sensing and machine learning algorithms: A review on progress, challenges, and opportunities." *Remote Sensing* 17.14 (2025).
- [13]. Wilhelm, Roland C., Harold M. van Es, and Daniel H. Buckley. "Predicting measures of soil health using the microbiome and supervised machine learning." *Soil Biology and Biochemistry* 164 (2022).
- [14]. Wilhelm, Roland C., Harold M. van Es, and Daniel H. Buckley. "Predicting measures of soil health using the microbiome and supervised machine learning." *Soil Biology and Biochemistry* 164 (2022).
- [15]. Ye, Xiao-Wei, Tao Jin, and Yun-Min Chen. "Machine learning-based forecasting of soil settlement induced by shield tunneling construction." *Tunnelling and Underground Space Technology* 124 (2022).
- [16]. Mishra, Dipanshu, et al. "A Review on Ensemble Learning Methods: Machine Learning Approach." *International Journal of Research Publication and Reviews* 6.2 (2025)