NJSS 35(2) 2026: 1-11 DOI: 10.67042/njss.2026.xxxxxxxxx

Nigerian Journal of Soil Science

Review Article

Digital Soil Mapping of Soil Taxonomic Classes: Prospects and Challenges for Machine Learning Integration in Nigeria

Yasin Agono Awwal[1,2*], Lemuel Musa Maniyunda[2], Sulaiman Lawan Ya'u[2], Muhammad Isma'il[3]

[1]Department of Soil Science and Land Resources Management, Federal University Wukari, Taraba State, Nigeria
[2]Department of Soil Science, Ahmadu Bello University, Zaria, Kaduna State, Nigeria
[3]Department of Geography and Environmental Management, Ahmadu Bello University, Zaria, Kaduna State, Nigeria

***Corresponding Author Details:**
Email: awwalyasin@fuwukari.edu.ng

Keywords:
digital soil mapping, machine learning, soil classification, soil information system, Nigeria

Manuscript History:
Submitted: 13th August 2026
Accepted: 01st September 2026
Published:

Abstract

Reliable spatially explicit soil information is important for land-use planning, agricultural management, and food-security planning in Nigeria. Digital soil mapping (DSM), which combines soil observations with environmental covariates through spatial prediction functions, provides an opportunity to complement and progressively update existing soil information. This review examines the prospects and challenges of integrating machine learning (ML) into DSM of soil taxonomic classes in Nigeria. Peer-reviewed literature was synthesized to assess soil classification systems, environmental covariates, modelling approaches, validation practices, and the relative development of continuous soil-property mapping and discrete taxonomic-class prediction. The evidence retrieved at the time of literature search indicates that ML-based DSM of continuous properties, such as soil organic carbon, particle-size fractions, and other nutrient indicators is more developed in Nigeria than direct prediction of formal soil taxonomic classes. This highlights an important methodological and information gap, constraining evidence on direct ML-based taxonomic-class prediction and limiting the availability of spatially explicit taxonomic information for land-use planning, soil management, suitability assessment, and broader soil-resource decision-making. Key challenges identified include sparse and spatially clustered legacy data, limited high-resolution covariate availability, inconsistent harmonization between classification systems, constrained computing and technical capacity, and poor model transferability across agro-ecological zones. Prospects for advancing the field include renewed legacy-data mobilization, cloud-based access to high-resolution remote-sensing covariates, hybrid and deep-learning approaches, and deliberate investment in taxonomic-class-specific training datasets. Future Nigerian DSM research may prioritize direct classification of soil taxonomic units using ML, alongside refinement of continuous-property mapping, to build a coherent, nationally consistent soil information system.

1. Introduction

Soil is the foundation of Nigeria's agrarian economy, supporting the livelihoods of a majority of the population and underpinning national food-security objectives (Awwal *et al.*, 2026a). The country spans a wide range of agro-ecological zones, from the semi-arid Sahel and Sudan savanna in the north, through the Guinea savanna and derived savanna of the Middle Belt, to the humid rainforest and mangrove/coastal zones of the south, each associated with distinctive soil-forming factors and, consequently, distinctive soil taxa (Odeh *et al.*, 2012). Accurate knowledge of the spatial distribution of these soils, expressed through formal classification systems such as the United States Department of Agriculture (USDA) Soil Taxonomy and the World Reference Base for Soil Resources (WRB), is essential for land evaluation, crop suitability assessment, fertiliser recommendation, and the design of sustainable land-management interventions, when integrated with other specific property measurements.

Nigeria's national soil information base derives largely from a reconnaissance-level soil survey conducted by the Federal Department of Agricultural Land Resources (FDALR) between 1985 and 1990, mapped at a scale of 1:650,000 (Nkwunonwo *et al.*, 2020). While historically valuable, this survey and the semi-detailed and detailed surveys that followed in specific localities are constrained by coarse mapping-unit resolution, inconsistent georeferencing, variable depth reporting, and reliance on manual, expert-driven delineation (Odeh *et al.*, 2012). These limitations have motivated a shift, mirrored globally, toward digital soil mapping (DSM): the use of field and legacy soil observations together with quantitative environmental covariates to predict soil taxonomic classes or properties across the landscape through statistically defined spatial prediction functions, formalized in the SCORPAN framework of McBratney *et al.* (2003), in which soil is modelled as a function of existing soil information (s), climate (c), organisms (o), relief (r), parent material (p), age (a) and spatial position (n).

Within this framework, machine learning (ML) algorithms such as Random Forest, Support Vector Machines, Artificial Neural Networks, Cubist, boosted regression trees, and gradient-boosting ensembles such as XGBoost and CatBoost have progressively favoured over simpler linear and geostatistical approaches, owing to their capacity to capture non-linear, high-order interactions between soil and environmental covariates (Hengl *et al*., 2015; Nenkam *et al*., 2024). In Nigeria and comparable West African settings, ML-based DSM has been applied with encouraging results to continuous soil properties such as organic carbon, particle-size fractions and macronutrients (Akpa *et al*., 2014, 2016; Forkuor *et al*., 2017). However, the extension of the same machine-learning toolkit to the direct prediction of discrete soil taxonomic classes has received comparatively little attention in the Nigerian literature, even though taxonomic-class mapping is methodologically well established elsewhere (Brungard *et al*., 2015; Duan *et al*., 2024; Cherlinka *et al*., 2025).

This review addresses that gap directly. Its objectives were to: (i) synthesise the classification systems concurrently in use for Nigerian soils and the challenges of reconciling them; (ii) review the machine-learning algorithms, environmental covariates, and validation approaches applied in Nigerian and closely comparable West African digital soil mapping studies; (iii) distinguish the state of the art in continuous soil-property mapping from the substantially less developed state of taxonomic-class mapping; and (iv) articulate the principal challenges constraining, and prospects for advancing, machine-learning-integrated soil classification in Nigeria.

2. Methodology

2.1 Literature Search

Relevant literature was identified through structured searches of Scopus, Web of Science and Google Scholar, supplemented by direct retrieval from publisher platforms (Elsevier ScienceDirect, Springer Nature, Wiley, MDPI, Frontiers, PLOS and Nature Portfolio). Searches were conducted in July 2026, with no restriction on publication year. Search terms combined soil-classification and digital-soil-mapping vocabulary (“soil classification”, “soil taxonomy”, “digital soil mapping”, “predictive soil mapping”, “soil taxonomic class”) with ML vocabulary (“machine learning”, “random forest”, “support vector machine”, “artificial neural network”, “gradient boosting”, “deep learning”) and geographic qualifiers (“Nigeria”, “West Africa”, “sub-Saharan Africa”, “tropical”). The principal search expression was: ("soil classification" OR "soil taxonomy" OR "soil taxonomic class" OR "soil class") AND ("digital soil mapping" OR "digital soil assessment" OR "predictive soil mapping") AND ("machine learning" OR "random forest" OR "support vector machine" OR "artificial neural network" OR "deep learning" OR "XGBoost" OR "LightGBM") AND ("Nigeria" OR "West Africa" OR "Africa" OR "sub-Saharan Africa" OR "tropical"). Database-specific adaptations were applied where required by individual search interfaces.

Because Nigeria-specific studies predicting discrete soil taxonomic classes using ML proved scarce, the search was deliberately widened to include: (a) Nigerian and West African studies that map continuous soil properties using comparable DSM workflows; (b) studies documenting correspondence between soil-classification systems in Nigerian soils; and (c) methodologically influential non-African taxonomic-class prediction studies that provide relevant methodological benchmarks.

2.2 Study Selection and Eligibility

Studies were included if they were peer-reviewed and addressed soil classification, DSM, ML-based soil prediction, or closely related quantitative soil-mapping methods, with sufficient methodological information for assessment. Nigerian studies were prioritized, while relevant West African and African comparator studies were included alongside global methodological benchmarks. Studies were excluded if they were duplicates, non-peer-reviewed materials without sufficient methodological information, or lacked identifiable soil data, environmental covariates, modelling procedures, or classification methods. The screening process involved title and abstract screening followed by full-text assessment, with duplicate records removed before screening.

3. Soil Classification Systems in Use in Nigeria

Nigerian soil-survey and soil-mapping literature draws on several, only partially reconcilable, classification systems, summarised in Table 1. USDA Soil Taxonomy, now in its thirteenth edition (Soil Survey Staff, 2022), is one of the most frequently applied system in pedon-level characterisation studies, typically reported to subgroup level. The WRB is used alongside USDA Taxonomy in most published characterisation studies for international comparability, and is the taxonomic label most often reported in Nigerian digital-soil-mapping-adjacent studies of continuous properties, such as Akpa *et al*. (2016), who reported soil organic carbon density by WRB reference soil group.

A legacy layer, the FAO/UNESCO Soil Map of the World's 1988 revised legend, underlies both older correlation tables and the original FDALR reconnaissance survey, later digitised into an open national database by Nkwunonwo *et al*. (2020). Finally, the FDALR soil-association–soil-series concept, a field-mapped but non-formally-taxonomic system tied to physiography, parent material and drainage, remains the structural basis of Nigeria's national reconnaissance map, and was identified by Odeh *et al*. (2012) as inconsistently georeferenced and too coarse to support contemporary digital soil mapping without substantial harmonisation.

A recurring methodological difficulty documented in the literature is the lack of a standardised crosswalk between these systems. A recent correspondence analysis of 142 harmonised soil profiles in northern Kaduna State applied contingency analysis, cross-classification and Random Forest variable-importance modelling to quantify the environmental

Table 1. Soil classification systems referenced in the Nigerian digital soil mapping and soil classification literature

System	Structural basis	Application in the Nigerian/tropical literature reviewed
USDA Soil Taxonomy (Soil Survey Staff, 2022)	Hierarchical: order, suborder, great group, subgroup, family, series, based on diagnostic horizons and soil moisture/temperature regimes	The most frequently applied system in Nigerian pedon-level characterisation studies; rarely the direct output variable of published Nigerian machine-learning models.
World Reference Base for Soil Resources – WRB (IUSS Working Group WRB, 2022)	Reference soil groups defined by diagnostic horizons, properties and materials, refined with prefix/suffix qualifiers	Used alongside USDA Taxonomy for international comparability; the reference-soil-group level is the taxonomic label most often reported in Nigerian property-mapping studies
FAO/UNESCO Soil Map of the World legend (1988 revised legend)	Precursor to WRB; broad soil units mapped at 1:5,000,000	Basis of older Nigerian correlation tables and of the original Federal Department of Agricultural Land Resources (FDALR) reconnaissance soil survey (1985–1990)
Local/reconnaissance mapping units (FDALR soil-association–soil-series concept)†	Field-mapped soil associations tied to physiography, parent material and drainage; not a formal taxonomic hierarchy	Underlies Nigeria's national reconnaissance soil map; identified by Odeh et al. (2012) as inconsistently georeferenced and coarse in scale, motivating renewed legacy-data harmonisation for digital soil mapping.

†Not a formal taxonomic hierarchy, however, its product has been used widely as source for soil property differentiation

correlates of taxonomic-class differentiation, reporting substantial but incomplete agreement between WRB reference soil groups and USDA suborders (Cramér's V = 0.69), with slope position the strongest single environmental correlate of class membership in both systems (WRB V = 0.66; USDA V = 0.58) (Awwal *et al*., 2026b). Such correspondence studies are valuable precisely because they quantify, rather than assume, the degree to which Nigerian soils classified under different systems can be treated as equivalent, which is an essential precondition for any machine-learning model that seeks to predict taxonomic class as a harmonized training label across data compiled from multiple legacy sources.

4. Conventional Soil Survey and its Limitations in Nigeria

Nigeria's conventional soil-survey record consists predominantly of reconnaissance-scale mapping (1:650,000), together with scattered semi-detailed and detailed surveys conducted at university, state and project level. Odeh *et al*. (2012), in one of the most consequential assessments of this record, catalogued approximately 1,220 legacy soil profiles and found that around 35% originated from reconnaissance-level surveys, with data richness diminishing markedly as survey scale coarsened; many profiles also lacked adequate georeferencing or consistent depth reporting relative to the two-meter depth specification required by contemporary global soil-mapping standards. Nkwunonwo *et al*. (2020) subsequently digitized the FDALR paper maps into an open, GIS-based national soil database covering texture, slope, drainage, depth and mechanized-farming suitability, providing a modern, freely accessible entry point to this legacy record. Nonetheless, both studies underscore that conventional survey products, however valuable as a data source, cannot by themselves support the resolution or currency of soil information increasingly demanded for precision agriculture, environmental monitoring and land-degradation assessment, hence the need for integrating digital soil mapping, and ML.

5. Digital Soil Mapping Framework: From SCORPAN to Machine Learning

Digital soil mapping formalizes the long-standing recognition that soil properties and classes are systematically related to observable environmental factors. The SCORPAN framework (McBratney *et al*., 2003) expresses this relationship as a spatial prediction function in which a soil attribute of interest is modelled from soil information (s), climate (c), organisms/vegetation (o), relief (r), parent material (p), age (a) and spatial position (n), plus an error term. Early implementations of this framework relied on multiple linear regression and geostatistical interpolation; more recent implementations, including the great majority of the studies reviewed here, instead use machine-learning algorithms to fit the s = f(SCORPAN) relationship, exploiting their capacity to model non-linear and interacting effects among covariates without requiring a pre-specified functional form (Hengl *et al*., 2015).

In practice, the digital-soil-mapping workflow applied across the reviewed studies follows a consistent sequence involving compilation and harmonization of legacy and/or newly collected soil observations to a common target legend; assembly of an environmental covariate stack; partitioning of data for model training and validation; model fitting and covariate-importance assessment; and, finally, generation of spatially continuous prediction maps together with an accompanying measure of uncertainty (Figure 1). Where the target is a continuous property, the model output is a regression surface; where the target is a taxonomic class, the output is a set of class-probability rasters from which the most probable class, and its associated uncertainty, can be mapped (Brungard *et al*., 2015).

6. Machine Learning Algorithms Applied in Soil Mapping

Random Forest is the most frequently applied algorithm across the reviewed literature, valued for its robustness to noisy and correlated covariates, its ability to handle mixed data types, and its built-in variable-importance diagnostics (Hengl *et al*., 2015; Akpa *et al*., 2014, 2016). Cubist and boosted regression trees have been used as comparators in Nigerian organic-carbon mapping, generally performing competitively but less consistently than Random Forest (Akpa *et al*., 2016). Support Vector Machines and Artificial Neural Networks appear in both Nigerian nutrient/texture-prediction studies (Folorunso *et al*., 2025) and in regional West African property-mapping comparisons (Forkuor *et al*., 2017), typically alongside Random Forest as part of a multi-model comparison rather than as the primary or sole algorithm. Gradient-boosting ensembles such as XGBoost, LightGBM and CatBoost have entered the Nigerian literature more recently, applied to organic-carbon mapping in northern Nigeria alongside Quantile Random Forest for uncertainty estimation (Abdulkadir *et al*., 2026), and to soil classification in the neighbouring Sudan savanna of Burkina Faso using Sentinel-1/2 imagery (Maung *et al*., 2026).

For the specific task of taxonomic-class prediction the most directly relevant methodological reference remains Brungard *et al*. (2015), a non-African but widely cited benchmark study that systematically compared clustering algorithms, discriminant analysis, multinomial logistic regression, neural networks, tree-based methods and support vector classifiers for predicting soil taxonomic class across three semi-arid U.S. landscapes. That study found that model complexity mattered: complex models, and Random Forest combined with covariates selected via recursive feature elimination in particular, consistently outperformed simpler classifiers. Given the strong performance of Random Forest across the Nigerian continuous-property literature as well, this consistency across both property- and class-prediction tasks, and across geographically disparate settings, is a reasonable basis for prioritizing Random Forest as the starting algorithm for future Nigerian taxonomic-class mapping efforts.

7. Applications of ML in Digital Soil Mapping Relevant to Nigeria

7.1 Continuous soil-property mapping

The best-developed strand of Nigerian ML-DSM literature addresses continuous soil properties. Akpa *et al*. (2014) predicted particle-size fractions across Nigeria at six GlobalSoilMap-standard depths using Random Forest, Cubist and regression kriging, finding that Random Forest showed the strongest overall performance among the evaluated approaches, with accuracy declining with depth (clay $R^2 = 0.53$ at 0–5 cm versus 0.16 at 100–200 cm). Building on this, Akpa *et al*. (2016) mapped total soil organic carbon and national sequestration potential, again finding Random Forest superior to Cubist and boosted regression trees in most cases, and estimating approximately 6.5 Pg of organic carbon stored in the top metre of Nigerian soils nationally, concentrated under forest vegetation and Ferralsols. More recently, a comparative study in northern Nigeria applied Random Forest, XGBoost and CatBoost to map organic carbon at 30 m resolution from a reduced covariate set (12 of 22 candidates), again finding Random Forest most accurate and using Quantile Random Forest to quantify prediction uncertainty (Abdulkadir *et al*., 2026). At the nutrient scale, Folorunso *et al*. (2025) applied an Artificial Neural Network to predict soil texture class from

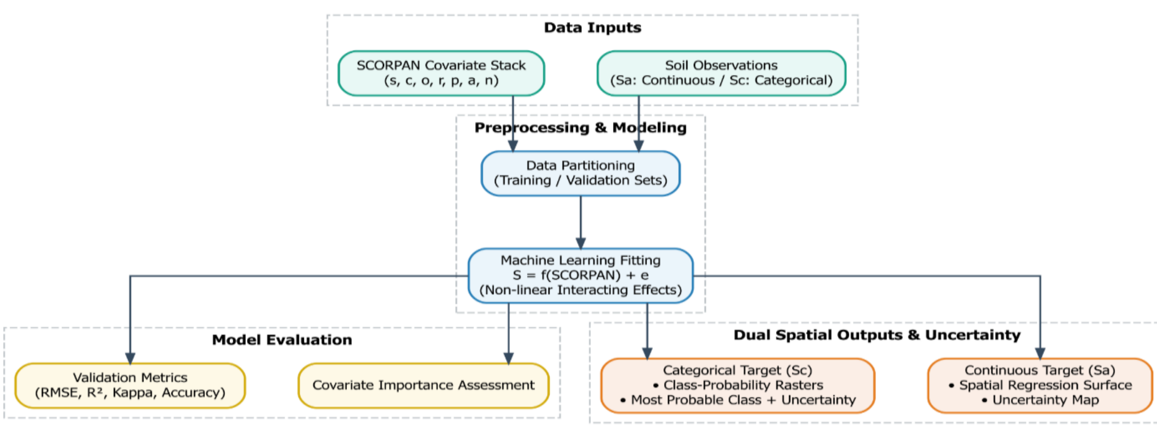

Figure 1. Typical Digital Soil Mapping Workflow Based on the McBratney *et al*. (2003) SCORPAN Framework

Because Nigeria-specific taxonomic-class studies remain limited, regional West African evidence is instructive. Forkuor *et al*. (2017) compared Random Forest, Support Vector Machines, Stochastic Gradient Boosting and multiple linear regression for mapping six soil properties in south-western Burkina Faso using high-resolution RapidEye and Landsat imagery, finding that all three machine-learning methods outperformed linear regression, and underscoring the value of remote-sensing covariates in data-scarce settings closely comparable to much of northern and Middle Belt Nigeria. More directly relevant to taxonomic-class prediction, a 2026 study in the Sudan savanna of Burkina Faso's Central Plateau used Sentinel-1 and Sentinel-2 spectral and topographic indices with XGBoost, Random Forest and Support Vector Machines to classify soil types directly, demonstrating the practical feasibility of satellite-based taxonomic classification in an agro-ecological zone that has a close analogue across much of northern Nigeria.

7.3 Studies bearing on taxonomic-class mapping

Three strands of evidence bear most directly on the taxonomic-classification component of this review's title. First, the Kaduna WRB–USDA correspondence study discussed in Section 3 used Random Forest-based variable-importance analysis to quantify environmental controls on taxonomic-class differentiation, offering a template for covariate selection in a future Nigerian class-prediction model. Second, Babatunde *et al*. (2026) provide a rigorously classified, morphology-based dataset for southern Ekiti State, assigning profiles to both USDA Soil Taxonomy and WRB (predominantly Ultisols) and linking soil morphology to agronomic potential for maize, cassava and Bologi; exactly the kind of harmonized, well-documented ground-truth dataset that a Nigerian machine-learning taxonomic-class model would require for training and validation. Third, the semi-arid U.S. benchmark of Brungard *et al*. (2015) remains the most complete methodological demonstration, within the wider global literature, of how covariate selection strategy and model complexity jointly determine taxonomic-class prediction accuracy, and offers a directly transferable experimental design for a Nigerian study.

7.4 Continental-scale context

Nigeria's position within continental efforts is documented by two studies. Hengl *et al*. (2015) incorporated Nigerian legacy profiles into a pan-African, 250 m resolution Random Forest ensemble mapping multiple continuous soil properties, demonstrating clear accuracy gains over earlier continental soil grids. Nenkam *et al*. (2024), reviewing digital soil mapping practice across Africa, found that roughly half of African countries have at least one published digital soil mapping study, but identified covariate scarcity, legacy-data positional and temporal error, and low, spatially clustered sampling density as persistent, continent-wide constraints. A related extrapolation study using the ISRIC Africa Soil Profiles database found that Random Forest models trained on data from one African country, including Nigeria, degrade sharply when applied to a neighbouring country, indicating that feature-space transferability across national or agro-ecological boundaries cannot be assumed and that within-country training data remain essential (Hateffard *et al*., 2024).

8. Environmental Covariates and Data Sources

Consistent with the SCORPAN framework, the covariate stacks used across the reviewed studies cluster into four broad categories. Climatic covariates, which are typically long-term rainfall and temperature surfaces used to represent the north–south gradient from Sahelian to humid-forest conditions that strongly structures Nigerian soil formation. Terrain covariates, derived principally from the Shuttle Radar Topography Mission (SRTM) digital elevation model, capture slope, curvature, wetness index and other relief attributes that Awwal *et al*. (2026b) found to be the strongest single predictor of taxonomic-class membership. Remote-sensing covariates, derived from Landsat and, increasingly, Sentinel-1/2 imagery, provide vegetation and spectral indices used as proxies for organic-matter content, land cover and, in the Sudan savanna classification study, direct inputs to taxonomic classification. Finally, legacy soil-profile data, usually from the FDALR reconnaissance survey, university and project-level surveys, or continental compilations such as the ISRIC Africa Soil Profiles database supply the training observations against which all of the above covariates are related, making the quality and harmonisation issues discussed in Sections 3 and 4 directly consequential for model performance.

9. Model Validation and Uncertainty Quantification

Continuous-property studies in the reviewed literature predominantly report the coefficient of determination (R2) and root mean square error (RMSE) under k-fold or holdout cross-validation; several studies additionally report the divergence between random and spatial cross-validation, with spatial cross-validation consistently yielding lower, and arguably more realistic, accuracy estimates because it better represents prediction at unsampled locations. Categorical taxonomic-class studies, exemplified by Brungard *et al*. (2015), instead report overall accuracy, kappa coefficients and class-specific producer's/user's accuracy derived from confusion matrices. Uncertainty quantification is increasingly reported alongside point accuracy: Quantile Random Forest has been used to map spatially explicit prediction uncertainty for organic carbon in northern Nigeria (Abdulkadir *et al*., 2026), while Dissimilarity Index and Area of Applicability metrics have been used elsewhere in the wider literature to delineate regions where

Table 2. Synthesis of machine-learning digital soil mapping studies relevant to soil classification and mapping in Nigeria

Study	Study area	ML algorithm(s)	Target output	Classification / legend	Headline finding
Odeh et al. (2012)	Nigeria (national, legacy data)	None (legacy-data audit)	Compilation/QC of ~1,220 legacy soil profiles	Mixed legacy legends	Poor georeferencing and coarse survey scale (≈35% reconnaissance-level) identified as key barriers to DSM readiness.
Akpa et al. (2014)	Nigeria (national)	Random Forest, Cubist, regression kriging	Particle-size fractions at 6 GlobalSoilMap depths	N/A (continuous)	Random Forest gave the most reliable predictions; accuracy declined with depth (clay R^2 = 0.53 at 0–5 cm vs 0.16 at 100–200 cm).
Akpa et al. (2016)	Nigeria (national)	Random Forest, Cubist, Boosted Regression Trees	Total soil organic carbon stock	Reported by WRB reference soil group	Random Forest outperformed the other two models in most cases; ≈6.5 Pg SOC estimated in the top 1 m nationally, highest under forest/Ferralsols.
Hengl et al. (2015)	Sub-Saharan Africa incl. Nigeria (250 m)	Random Forest ensemble	Multiple soil properties (OC, pH, particle size, CEC, TN, bases)	N/A (continuous)	Random-forest ensembling significantly improved accuracy over earlier continental soil grids.
Forkuor et al. (2017)	South-western Burkina Faso	Random Forest, SVM, Stochastic Gradient Boosting, MLR	Sand, silt, clay, CEC, SOC, nitrogen	N/A (continuous)	Machine-learning models outperformed multiple linear regression; remote-sensing covariates improved local-scale DSM in a data-scarce West African setting.
Nkwunonwo et al. (2020)	Nigeria (national)	None (GIS digitisation)	Digitised national soil map/database (texture, slope, drainage, depth)	Legacy FDALR reconnaissance classes	Produced a free, open national digital soil database from 1985–1990 paper maps — a candidate legacy source for future ML taxonomic-class mapping.
Nenkam et al. (2024)	Africa-wide, incl. Nigeria	Review of RF, SVM, ANN, boosted trees	Continental review of DSM practice	Mixed (WRB, USDA, local)	Nearly half of African countries have at least one DSM study; covariate scarcity, legacy-data error and low, clustered sampling density are recurring constraints.
Brungard et al. (2015)	Semi-arid USA (methodological benchmark)	Random Forest, SVM, neural networks, multinomial logistic regression, discriminant analysis	Soil taxonomic class (order/subgroup)	USDA Soil Taxonomy	Complex models — especially Random Forest with covariates chosen by recursive feature elimination — gave the most accurate taxonomic-class predictions; a key methodological reference for tropical class-level DSM.

Table 2 Continued.

Study	Study area	ML algorithm(s)	Target output	Classification / legend	Headline finding
Abdulkadir *et al.* (2026)	Northern Nigeria (agricultural land)	Random Forest, XGBoost, CatBoost; Quantile Random Forest	Soil organic carbon at 30 m resolution	N/A (continuous)	Random Forest was most accurate (R^2 = 0.43); Quantile Random Forest used to map prediction uncertainty.
Maung *et al*. (2026)	Central Plateau, Burkina Faso	Random Forest, XGBoost, SVM	Soil class prediction	Locally defined classes	Demonstrates feasibility of Sentinel-based ML soil classification in a data-scarce Sudano-Sahelian setting directly analogous to northern Nigeria.
Awwal *et al*. (2026b)	Northern Kaduna State, Nigeria	Random Forest (variable importance/contingency analysis)	Correspondence between WRB reference soil groups and USDA suborders	WRB (2022) and USDA Soil Taxonomy (2022)	Slope position showed the strongest environmental association with soil class (WRB Cramérs V = 0.66; USDA V = 0.58); WRB–USDA cross-classification agreement was substantial (Cramér's V = 0.69).
Babatunde *et al*. (2026)	Southern Ekiti State, Nigeria	None (conventional morphological classification)	Soil profile morphology and land-suitability evaluation	USDA Soil Taxonomy and WRB (Ultisols)	Links tropical soil-profile morphology to agronomic potential; a candidate ground-truth benchmark for future taxonomic-class ML studies in south-western Nigeria.
Folorunso *et al.* (2025)	South-west Nigeria (6 states)	Artificial Neural Network	Soil texture class prediction	Textural class	ANN-predicted texture classes validated against iSDAsoil, feeding a farmer-facing fertiliser-recommendation tool.
Hateffard *et al.* (2024)	Ethiopia, Kenya, Burkina Faso, Nigeria	Random Forest	Topsoil organic carbon, clay content, pH	N/A (continuous)	Random Forest models trained in one African country degrade sharply when extrapolated to another, underlining limited transferability between Nigeria and neighbouring countries.

model predictions can be trusted versus regions requiring extrapolation. We recommend that future Nigerian taxonomic-class studies report both class-probability rasters and an explicit measure of prediction confidence, given the practical decision-making consequences of taxonomic misclassification for land-use recommendations.

An additional consideration in DSM is soil depth, which can substantially influence predictive performance and has important implications for taxonomic soil classification. Soil properties do not exhibit the same degree of predictability at all depths because their relationships with environmental covariates and landscape processes vary vertically within the soil profile. Studies that evaluate predictions separately by depth have consequently reported differences in model performance between surface and subsurface layers (Figure 2; Akpa *et al*., 2014; 2016). Their findings indicate that model performance, measured as R2, decreased from 0.53 – 0.48 in surface 0 – 5 cm depth, to 0.13 at deeper 100 – 200 cm layers (Figure 2). This depth dependence suggests that many environmental covariates used in DSM researches provide information that are more directly related to surface conditions than to properties occurring deeper in the profile.

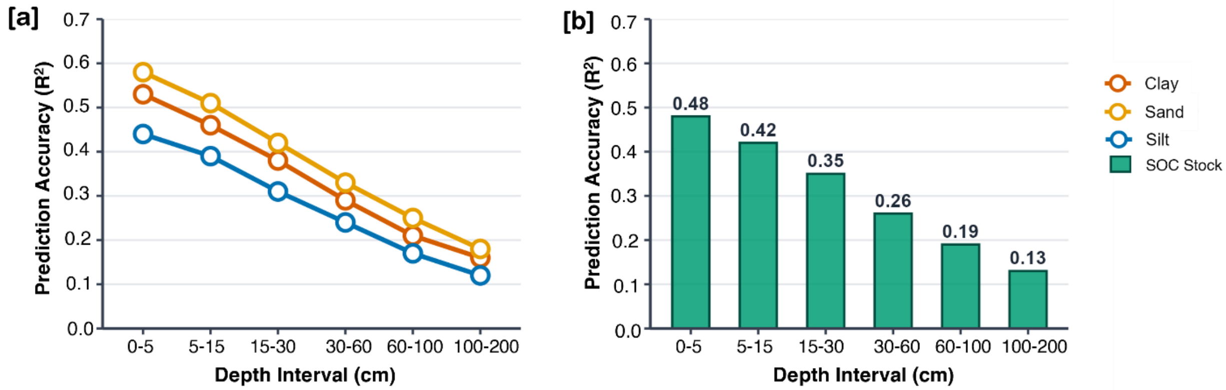

Figure 2. Depth Attenuation of Machine Learning Model Accuracy in Nigerian Digital Soil Mapping. Model performance across standard GlobalSoilMap target depths based on Random Forest predictions of [a] Particle-Size Fractions [b] Soil Organic Carbon Stock. Figures depict model accuracy reported for varying properties predicted at different depths.

For soil taxonomic classification, however, the relationship is more complex. Surface horizons and their properties may provide useful indirect indicators of the environmental conditions under which a soil developed, and may therefore contribute substantially to discrimination among some taxonomic classes. At the same time, many taxonomic distinctions depend on diagnostic subsurface horizons or properties, such as clay accumulation, plinthite occurrence, redoximorphic features, or other diagnostic characteristics that may occur below the surface. Consequently, strong predictive performance for surface soil properties should not automatically be interpreted as evidence that an ML model can reliably reproduce taxonomic classes whose defining criteria occur at depth.

The findings of Akpa *et al*. (2014; 2016), which demonstrate differences in predictive performance across depth intervals, are therefore particularly relevant to taxonomic DSM. They suggest that the relationship between environmental covariates and soil information is depth-dependent and that model performance should be evaluated separately across relevant soil-depth intervals rather than assuming that a model developed from surface observations will perform equivalently throughout the profile. This consideration is especially important when taxonomic classification is inferred indirectly from predicted soil properties. For example, where a taxonomic class is distinguished by subsurface clay accumulation, drainage conditions, or other diagnostic properties, prediction models based only on surface observations may capture the environmental setting of the soil without adequately representing the diagnostic property itself. Where sufficient observations are available, studies should examine whether predictor importance and model performance change systematically with depth and should identify the depth intervals most informative for the target taxonomic units. This would help distinguish between models that predict surface environmental conditions effectively and models that genuinely capture the vertical soil characteristics required for taxonomic classification.

10. Challenges of ML Integration for Soil Classification in Nigeria

The major challenges affecting machine-learning-based soil classification in Nigeria extend from limitations in legacy soil data and sampling design to taxonomic inconsistency, class imbalance, incomplete profile information, covariate limitations, and environmental distribution shifts. These constraints have consequences not only for pedological interpretation but also for model training, validation, prediction accuracy and transferability. Table 3 synthesizes these challenges by linking their pedological consequences and implications for ML performance with practical solutions that can strengthen future taxonomic-classification studies in Nigeria, while more explanations trail.

10.1 Legacy data quality and sampling density

The foundational challenge, documented since Odeh *et al*. (2012) and echoed by Nenkam *et al*. (2024) at continental scale, is the sparse, spatially clustered and inconsistently georeferenced nature of Nigeria's legacy soil-profile record, which constrains both the training sample size and the spatial representativeness available to any machine-learning model.

10.2 Covariate availability and resolution

High-resolution terrain, climate and remote-sensing covariates are unevenly available across Nigeria, while limitations in detailed parent-material and geological mapping constrain the covariate information available for differentiating soil taxonomic classes. This limitation is particularly important because parent material influences mineral composition, weathering pathways, texture, drainage characteristics and other pedogenic properties that contribute to soil differentiation (Fatihu *et al*., 2020). In Nigeria, geological information is often derived from relatively coarse-scale regional geological maps, which may generalize complex lithological units and therefore provide limited discrimination at the scale required for soil mapping.

Table 3. Key Challenges, Pedological and Machine-Learning Consequences, and Recommended Solutions for ML-Based Soil Classification in Nigeria

Challenge	Pedological consequences	ML consequence	Recommended solution
Sparse legacy data	Poor representation of classes	Overfitting/extrapolation	Targeted sampling
Spatial clustering	Sampling bias	Optimistic CV	Spatial CV
Taxonomic inconsistency	Label noise	Poor classification	Harmonization
Rare taxa	Class imbalance	Poor minority-class recall	Stratified sampling/class weighting
Missing depth information	Loss of diagnostic horizons	Misclassification	Profile-level sampling
Covariate mismatch	Weak soil-environment relationship	Lowe predictive power	Multi-source covariates
Agroecological transfer	Distribution shift	Poor transferability	AOA/domain adaptation

This is further complicated by heterogeneous lithology within mapped geological units and by the occurrence of transported materials, including alluvial and colluvial deposits, that may obscure the relationship between surface geology and the material from which soils actually develop. Prolonged weathering and pedogenic transformation can further modify the original mineralogical and chemical signatures of parent materials, while erosion, deposition and landscape redistribution may introduce materials from upslope or upstream sources. Consequently, geological boundaries do not necessarily correspond directly with soil boundaries, and uncertainty in mapped lithological contacts can propagate into soil-class predictions. These challenges are particularly relevant to taxonomic classification, where relatively subtle differences in diagnostic horizons and properties may distinguish classes occurring under similar climatic and topographic conditions (Awwal *et al*., 2026a). Integrating geology with terrain, remote-sensing and climatic covariates can therefore provide useful complementary information, but the predictive contribution of geological variables should be interpreted in light of their spatial scale, positional uncertainty and imperfect representation of the actual soil parent material.

10.3 Classification-system harmonization

As Section 3 details, Nigerian soil records are distributed across USDA Taxonomy, WRB, the legacy FAO/UNESCO legend and local reconnaissance units, with only partial, imperfectly quantified correspondence between them. Any machine-learning model that pools training data across sources must first resolve this harmonization problem, and the scarcity of correspondence studies such Awwal *et al*. (2026) is itself a barrier to progress.

10.4 Scarcity of taxonomic-class-specific training data

The literature synthesized in Section 7 shows a clear asymmetry: continuous-property mapping (SOC, particle size, nutrients) is comparatively mature, whereas machine-learning models trained explicitly to predict discrete taxonomic classes remain rare in Nigeria. This reflects both the harmonization problem above and the greater labour required to compile a taxonomically labelled, spatially representative training set relative to a continuous laboratory measurement.

10.5 Technical capacity and computing infrastructure

Machine-learning-based DSM requires familiarity with statistical programming environments, covariate-processing pipelines and, increasingly, cloud-based geospatial platforms; capacity in these skills, and access to the computing infrastructure they require, remains unevenly distributed across Nigerian research and government institutions relative to the international consortia (e.g., ISRIC, AfSIS) that have driven much of the continental-scale work reviewed here.

10.6 Model transferability

The extrapolation-potential study reviewed in Section 7.4 demonstrates that Random Forest models trained in one country, including Nigeria, generalize poorly to neighboring countries; by extension, models trained in one Nigerian agro-ecological zone should not be assumed to transfer reliably to another without explicit testing, a caution particularly relevant given Nigeria's marked north–south environmental gradient (Hateffard *et al*., 2024).

11. Prospects for Machine-Learning-Integrated Soil Classification in Nigeria

11.1 Legacy-data mobilization and national soil information systems

The digitization of the FDALR reconnaissance survey by Nkwunonwo *et al*. (2020) demonstrates that Nigeria's legacy soil archive can be rendered into an open, GIS-ready format. However, substantially greater value could be obtained by extending such efforts beyond map digitization to the development of a national soil-profile geodatabase incorporating the more detailed university- and project-level surveys catalogued by Odeh *et al*. (2012). Such a database should provide standardized information on soil taxonomy, morphology, analytical methods, geographic coordinates, depth intervals, and uncertainty metadata, together with consistent georeferencing and harmonised depth intervals. Establishing these standards would facilitate integration of heterogeneous legacy datasets while preserving information on data quality and provenance. A nationally coordinated, standardized soil-profile database would substantially expand the quality and quantity of training data available for taxonomic-class modelling and provide a stronger foundation for reproducible digital soil mapping across Nigeria.

11.2 High-resolution remote sensing and cloud computing

The demonstrated feasibility of Sentinel-1/2-based soil classification in the neighboring Sudan savanna of Burkina Faso (Maung *et al*., 2026), combined with cloud-based processing platforms that lower the computational barrier to assembling large covariate stacks, offers a potentially transferable pathway for Nigerian taxonomic-class mapping, particularly across the country's northern agro-ecological zones.

11.3 Hybrid, ensemble and deep-learning approaches

Random Forest has demonstrated strong and consistent performance across several continuous soil-property and taxonomic-classification studies, making it an important benchmark for digital soil mapping. However, the available evidence does not support assuming its superiority a priori. Recent Nigerian studies have reported competitive or superior performance from gradient-boosting algorithms, including XGBoost, for soil classification and soil-property prediction (Maung *et al*., 2026; Abdulkadir *et al*., 2026). Similarly, evidence from outside Africa indicates that ensemble performance can vary with the target soil property or taxonomic level, predictor set, sample characteristics and model configuration, with systematic covariate selection potentially providing additional gains (Brungard *et al*., 2015). Thus, no single ML algorithm should be assumed a priori to be optimal for Nigerian taxonomic-class mapping. Instead, comparative evaluation of multiple algorithms under consistent data, covariate and validation frameworks is necessary to identify the most appropriate model for a given Nigerian soil-mapping context.

11.4 Direct taxonomic-class prediction as a research priority

Given the clear gap identified in Section 7, we regard the direct machine-learning prediction of WRB reference soil groups and/or USDA great groups/subgroups rather than continuous properties alone as a high priority research direction for advancing Nigerian digital soil mapping. Well-documented, morphologically classified datasets such as that of Babatunde *et al*. (2026), combined with the covariate-selection and correspondence-analysis approaches demonstrated in the Kaduna study, provide a practical starting template.

11.5 Capacity building and institutional collaboration

Continued and expanded collaboration with continental initiatives such as AfSIS and ISRIC, which underpin the pan-African covariate stacks and legacy-profile compilations used by Hengl *et al*. (2015) and the extrapolation-potential study alongside targeted training in machine-learning workflows for Nigerian soil scientists, would help close the capacity gap identified.

11.6 Policy and stakeholder uptake

Ultimately, the value of improved taxonomic-class maps depends on their integration into land-use planning, fertilizer-recommendation, and national soil-monitoring frameworks; deliberate engagement with agricultural extension services and land-use policy bodies should accompany, rather than follow, the technical development recommended above.

12. Conclusion

This review has synthesized the peer-reviewed evidence on machine-learning-integrated digital soil mapping and soil classification in Nigeria. It shows that Nigerian and closely comparable West African studies have made substantial progress in applying machine-learning algorithms to map continuous soil properties such as organic carbon, particle-size fractions and macronutrients, often outperforming linear-regression and simpler statistical alternatives. By contrast, the direct machine-learning prediction of discrete soil taxonomic classes, whether under USDA Soil Taxonomy or the World Reference Base, remains comparatively underdeveloped in the Nigerian literature, with most taxonomic information still generated through conventional, non-spatial pedon classification. This asymmetry, together with persistent challenges of legacy-data quality, covariate availability, classification-system harmonization, technical capacity and model transferability, defines the principal agenda for future research. Addressing these challenges through legacy-data mobilization, expanded access to high-quality environmental covariates, methodologically rigorous taxonomic-class prediction studies informed by established benchmarks, and sustained institutional collaboration could complement and progressively update Nigeria's legacy soil information base, contributing toward a more current, quantitative, and taxonomically explicit national soil information system.

References

Abdulkadir, M., Csorba, A., Fuchs, M., Szegi, T., Hammani, B., Micheli, E., & Gelsleichter, Y. A. (2026). High resolution mapping of soil organic carbon in northern Nigerian agricultural lands using comparative machine learning with uncertainty analysis. *Discover Sustainability*, *7*, Article 1161. https://doi.org/10.1007/s43621-026-03407-2

Akpa, S. I. C., Odeh, I. O. A., Bishop, T. F. A., & Hartemink, A. E. (2014). Digital mapping of soil particle-size fractions for Nigeria. *Soil Science Society of America Journal*, *78*(6), 1953–1966. https://doi.org/10.2136/sssaj2014.05.0202

Akpa, S. I. C., Odeh, I. O. A., Bishop, T. F. A., Hartemink, A. E., & Amapu, I. Y. (2016). Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma*, *271*, 202–215. https://doi.org/10.1016/j.geoderma.2016.02.021

Awwal, Y. A., Maniyunda, L. M., Ya'u, S. L., Shobayo, A. B., & Gisilanbe, S. A. (2026b). Do tropical soil taxonomic classes reflect soil-forming factors? Evidence from soil profile observations and environmental covariates. *Geoderma Regional*, *45*, Article e01096. https://doi.org/10.1016/j.geodrs.2026.e01096

Awwal, Y. A., Sadiq, F. K., Gisilanbe, S. A., & David, J. E. (2026a). Characterization and Classification of Selected Soils Formed on a Planar Surface in Wukari, Nigeria. *Nigerian Journal of Soil Science*, 35(1), 239 - 256. https://doi.org/10.67042/njss.2026.en5g3nqd

Babatunde, C. A., Aruleba, R., Kumar, R., Kumar, P., Shahi, U. P., & Rajput, V. D. (2026). Soil profile morphology as an indicator of geo-environmental processes and agronomic potential in tropical landscapes of southern Ekiti State, Nigeria. *Frontiers in Soil Science*, *6*, Article 1920550. https://doi.org/10.3389/fsoil.2026.1920550

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, *239–240*, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019

Cherlinka, V., Gallay, M., Dmytruk, Y., & Dent, D. (2025). Predictive soil map of Slovakia: a digital twin for decision makers. *International Journal of Environmental Studies*, *82*(5–6), 1903–1925. https://doi.org/10.1080/00207233.2025.2539654

Duan, M., Song, X., Li, Z., Zhang, X., Ding, X., & Cui, D. (2024). Identifying soil groups and selecting a high-accuracy classification method based on multi-textural features with optimal window sizes using remote sensing images. *Ecological Informatics*, 81, 102563. https://doi.org/10.1016/j.ecoinf.2024.102563

Fatihu, K. S., Maniyunda, L. M. and Awwal, Y.A. (2020). Assessment of Soil Properties under different Land Uses and their Suitability for Maize (Zea mays) Production in Hunkuyi, Northern Guinea Savanna, Nigeria. *Dutse Journal of Agriculture and Food Security,* Vol. 7(1): 9-22 pp.

Folorunso, O., Ojo, O., Busari, M., Adebayo, M., Adejumobi, J., Folorunso, D., Ayo, F., Alabi, O., & Olabanjo, O. (2025). GeaGrow: A mobile tool for soil nutrient prediction and fertilizer optimization using artificial neural networks. *Frontiers in Sustainable Food Systems*, *9*, Article 1533423. https://doi.org/10.3389/fsufs.2025.1533423

Forkuor, G., Hounkpatin, O. K. L., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models. *PLOS ONE*, *12*(1), Article e0170478. https://doi.org/10.1371/journal.pone.0170478

Hateffard, F., Steinbuch, L., & Heuvelink, G. B. M. (2024). Evaluating the extrapolation potential of random forest digital soil mapping. *Geoderma*, *441*, Article 116740. https://doi.org/10.1016/j.geoderma.2023.116740

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Mendes de Jesus, J., Tamene, L., & Tondoh, J. E. (2015). Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PLOS ONE*, *10*(6), Article e0125814. https://doi.org/10.1371/journal.pone.0125814

IUSS Working Group WRB. (2022). World Reference Base for Soil Resources, international soil classification system for naming soils and creating legends for soil maps (4th ed., update 2022). International Union of Soil Sciences, Vienna.

Maung, W. S., Ikazaki, K., Sakai, T., Simporé, S., Zerbo, L., & Kone, N. (2026). Soil classification in the Sudan Savanna using sentinel products and topographic information with machine learning models. *Scientific Reports*, *16*, Article 10880. https://doi.org/10.1038/s41598-026-46259-6

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1–2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Nenkam, A. M., Wadoux, A. M. J.-C., Minasny, B., Silatsa, F. B. T., Yemefack, M., Ugbaje, S. U., Akpa, S., Van Zijl, G., Bouasria, A., Bouslihim, Y., Chabala, L. M., Ali, A., & McBratney, A. B. (2024). Applications and challenges of digital soil mapping in Africa. *Geoderma*, *449*, Article 117007. https://doi.org/10.1016/j.geoderma.2024.117007

Nkwunonwo, U. C., Okeke, F. I., Ebinne, E. S., & Chiemelu, N. E. (2020). Free, open, quantitative and adaptable digital soil map data and database for Nigeria. *Data in Brief*, *31*, Article 105941. https://doi.org/10.1016/j.dib.2020.105941

Odeh, I. O. A., Leenaars, J. G. B., Hartemink, A. E., & Amapu, I. (2012). The challenges of collating legacy data for digital mapping of Nigerian soils. In B. Minasny, B. P. Malone, & A. B. McBratney (Eds.), *Digital soil assessments and beyond* (pp. 453–458). CRC Press / Taylor & Francis. https://doi.org/10.1201/b12728-83

Soil Survey Staff. (2022). Keys to soil taxonomy (13th ed.). United States Department of Agriculture, Natural Resources Conservation Service, Washington, DC.