

الاء جاویش

Thesis final.docx

 Midocean University

Document Details

Submission ID

trn:oid:::2945:406744699

Submission Date

Aug 11, 2026, 8:15 PM GMT+3

Download Date

Aug 14, 2026, 5:27 PM GMT+3

File Name

Thesis final.docx

File Size

4.9 MB

39 Pages

8,518 Words

50,383 Characters

24% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

-  **178 Not Cited or Quoted 21%**
Matches with neither in-text citation nor quotation marks
-  **18 Missing Quotations 2%**
Matches that are still very similar to source material
-  **3 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 14%  Internet sources
- 19%  Publications
- 13%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 178 Not Cited or Quoted 21%**
Matches with neither in-text citation nor quotation marks
- 18 Missing Quotations 2%**
Matches that are still very similar to source material
- 3 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 14% Internet sources
- 19% Publications
- 13% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.mdpi.com	2%
2	Internet	www.researchgate.net	<1%
3	Publication	"Practical Statistical Learning and Data Science Methods", Springer Science and B...	<1%
4	Internet	etda.libraries.psu.edu	<1%
5	Internet	mdpi-res.com	<1%
6	Submitted works	Midocean University on 2025-09-05	<1%
7	Publication	"Sustainable Soil Systems in Global South", Springer Science and Business Media ...	<1%
8	Publication	Most. Mira Khatun, Md. Abir Hasan, Md Sadiul Alam Chyon, Rashid Anzoom et al. ...	<1%
9	Internet	link.springer.com	<1%
10	Publication	Cenk Temizel, Salih Tutun, Celal Hakan Canbaz, Ekrem Alagoz et al. "Artificial Inte...	<1%

11	Publication	Al Makdi, Khalid. "Adaptive Trust-Based Security Model for Intrusion Detection Us...	<1%
12	Publication	Hitesh Vasudev. "Artificial Intelligence in Tribology", CRC Press, 2026	<1%
13	Internet	www.avocadosource.com	<1%
14	Publication	Abhilasha Sharma, Vishwas Rathi, Anupam Biswas, Anil Singh, Omer Rana. "Multi...	<1%
15	Publication	Eva Corral-Pazos-de-Provens, Ígor Rapp-Arrarás, Juan M. Domingo-Santos. "Estim...	<1%
16	Publication	Mohamed A. E. AbdelRahman. "Remote sensing, geographic information system, ...	<1%
17	Publication	RV. Raghavendra Rao, U. Srinivasulu Reddy. "Agrimind intelligent deep learning f...	<1%
18	Internet	pure.tue.nl	<1%
19	Internet	vectoral.org	<1%
20	Submitted works	Middlesex University on 2024-04-28	<1%
21	Publication	"Producing Healthy Food with Healthy Soils", Springer Science and Business Medi...	<1%
22	Publication	Zhaoyu Zhai, Fuji Chen, Hongfeng Yu, Jun Hu, Xinfei Zhou, Huanliang Xu. "PS-MTL-...	<1%
23	Internet	media.neliti.com	<1%
24	Internet	unstats.un.org	<1%

25	Internet	www.udsspace.uds.edu.gh	<1%
26	Submitted works	The University of Manchester on 2025-03-24	<1%
27	Internet	theses.lib.polyu.edu.hk	<1%
28	Internet	doaj.org	<1%
29	Internet	pub.epsilon.slu.se	<1%
30	Internet	www.ijert.org	<1%
31	Publication	Bayazitova, Guzel. "Oil and Gas Flow Anomaly Detection on Offshore Naturally FL..."	<1%
32	Submitted works	University of Essex on 2003-11-06	<1%
33	Internet	core.ac.uk	<1%
34	Internet	drepo.sdl.edu.sa	<1%
35	Publication	Nadir El Bouanani, Ahmed Laamrani, Hicham Hajji, Mohamed Bourriz et al. "Esti..."	<1%
36	Submitted works	University of Western Australia on 2019-04-30	<1%
37	Publication	Zahra Ghahremani, David Huber, Arash Modaresi Rad, Mike J. Johnson, Jennifer Pi...	<1%
38	Publication	Abdulaziz Almohimeed, Redhwan M. A. Saad, Sherif Mostafa, Nora El-Rashidy et a...	<1%

39	Submitted works	Curtin University of Technology on 2024-01-24	<1%
40	Submitted works	Heriot-Watt University on 2025-08-07	<1%
41	Publication	Nathaniel W. Chaney, Budiman Minasny, Jonathan D. Herman, Travis W. Nauman ...	<1%
42	Publication	Osayande Pascal Omondiagbe, Linda Lilburne, Sherlock A. Licorish, Stephen G. M...	<1%
43	Internet	respiratory-research.biomedcentral.com	<1%
44	Publication	Ashutosh Sarangi, Sailesh Kumar Raula, Sohamdev Ghoshal, Swadhin Kumar, Chi...	<1%
45	Submitted works	Glyndwr University on 2025-12-17	<1%
46	Publication	M. Venkatanaresh, I. Kullayamma. "A Hybrid Cost-Sensitive and Feature Reweight...	<1%
47	Publication	T.S. Breure, A. Jones, P. Panagos. "Evaluating visible near-infrared spectroscopy in...	<1%
48	Internet	archive.saulibrary.edu.bd:8080	<1%
49	Internet	atxholiday.austintexas.org	<1%
50	Publication	Brigitta Szabó, Gábor Szatmári, Katalin Takács, Annamária Laborczi, András Makó...	<1%
51	Publication	Pratik Bhagyawant, Rutik Bansode, Swastik Ahiwale, Pradip P. Ghorpade. "AI-Bas...	<1%
52	Publication	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ...	<1%

53	Submitted works	University of Manouba on 2025-11-11	<1%
54	Publication	"Soil Health Series: Volume 4 Soil Health and Sustainability in India", Wiley, 2025	<1%
55	Publication	Dorijan Radočaj, Mateo Gašparović, Petra Radočaj, Mladen Jurišić. "Geospatial pr...	<1%
56	Submitted works	University of Sheffield on 2008-11-11	<1%
57	Publication	Zahia Aouabed, Vincent Therrien, Mohamed Achraf Bouaoune, Mohammadreza B...	<1%
58	Internet	aast.edu	<1%
59	Internet	hal.archives-ouvertes.fr	<1%
60	Internet	nur.nu.edu.kz	<1%
61	Publication	Badreddine Benmekhlouf, Cherif Ould-Lahoucine. "Dynamic thermodynamic mod...	<1%
62	Publication	Budhaditya Bhattacharyya, Anindita Kundu, Anantha Krishna Chintanpalli, Xavier...	<1%
63	Publication	Slim Smaoui, Angelos Vakalopoulos, Varzakas Theodoros. "Advances in Sustainab...	<1%
64	Publication	Wenchao Qi, Le Yu, Weichao Sun, Tao Liu, Qiang Zhao, Yue Zhao, Hui Wu, Yibo Wa...	<1%
65	Internet	cdn.techscience.cn	<1%
66	Internet	discovery.researcher.life	<1%

67	Internet	experts.umn.edu	<1%
68	Internet	file.techscience.com	<1%
69	Internet	netlibrary.aau.at	<1%
70	Internet	www.cognitivemarketresearch.com	<1%
71	Internet	www.researchsquare.com	<1%
72	Internet	www.ziprecruiter.com	<1%
73	Publication	"Computational Intelligence in Internet of Agricultural Things", Springer Science ...	<1%
74	Publication	"Proceedings of International Conference on Artificial Intelligence and Networks...	<1%
75	Publication	Abejide Ade-Ibijola, Solomon Oyelere, Jude Osakwe. "Innovation in Africa for Indu...	<1%
76	Submitted works	Cranfield University on 2008-10-07	<1%
77	Publication	Idowu, Ifedotun Roseline. "Improved Meta-Heuristic Based RNS Techniques for In...	<1%
78	Publication	Kathrin J. Ward, Sabine Chabrilat, Carsten Neumann, Saskia Foerster. "A remote ...	<1%
79	Submitted works	Nalla Malla Reddy Engineering College on 2026-02-06	<1%
80	Publication	Omid Abdi, Ville Laamanen, Jori Uusitalo. "Mapping forest fine-grained soil particl...	<1%

81	Publication	Rukhsana Sarkar, Asrafal Alam, Azizur Rahman Siddiqui. "Agriculture and Clima...	<1%
82	Submitted works	University of Sheffield on 2023-09-07	<1%
83	Submitted works	University of Sheffield on 2024-09-10	<1%
84	Internet	agridi.org	<1%
85	Internet	download.bibis.ir	<1%
86	Internet	etd.astu.edu.et	<1%
87	Internet	indah.ump.edu.my	<1%
88	Internet	rrs.mmu.edu.my	<1%
89	Submitted works	Anna University on 2024-08-09	<1%
90	Submitted works	Anna University on 2025-09-09	<1%
91	Submitted works	Cavite State University on 2024-01-27	<1%
92	Submitted works	Coventry University on 2024-09-11	<1%
93	Publication	Gaurav Aggarwal, Rishabh Dev Shukla, Innocent Ewean Davidson, Ramesh Bansal...	<1%
94	Submitted works	Imam Abdulrahman Bin Faisal University on 2025-12-17	<1%

95	Publication	Stella Mary Venkateswara, Jayashree Padmanaban. "Interpretable deep learning ..."	<1%
96	Submitted works	University of Edinburgh on 2024-02-29	<1%
97	Submitted works	University of Leicester on 2024-05-02	<1%
98	Submitted works	VIT University on 2025-04-09	<1%
99	Internet	arxiv.org	<1%
100	Internet	assets-eu.researchsquare.com	<1%
101	Internet	cendikiainovasi.org	<1%
102	Internet	experts.boisestate.edu	<1%
103	Internet	geoscience.ac	<1%
104	Internet	www.coursehero.com	<1%
105	Internet	www.epeu-umons.be	<1%
106	Internet	www.frontiersin.org	<1%
107	Internet	www.ijmrset.com	<1%
108	Internet	www.preprints.org	<1%

109	Publication	Arvind Dagur, Dharendra Kumar Shukla, Nazarov Fayzullo Makhmadiyarovich, Ak...	<1%
110	Publication	Bhabani Prasanna Pattanaik, Dharendra Kumar Shukla, Sambit Satpathy, Kamala...	<1%
111	Submitted works	Egerton University on 2026-01-23	<1%
112	Publication	Flavio Piccoli, Micol Rossini, Roberto Colombo, Raimondo Schettini, Paolo Napolet...	<1%
113	Publication	Giannis Gallios, Nikolaos Tsakiridis, Nikolaos Tziolas. "Federated learning applicat...	<1%
114	Submitted works	Higher Education Commission Pakistan on 2024-09-10	<1%
115	Publication	Leonardo Arrighi, Ingrid Alves de Moraes, Marco Zullich, Michele Simonato, Doug...	<1%
116	Publication	Mata, Abonga. "Evaluating the Use of Mid Infrared Spectroscopy in Predicting Soi...	<1%
117	Submitted works	University of Sydney on 2023-06-23	<1%
118	Publication	Achmad Arif Alfin, Nanik Suciati, Wijayanti Nurul Khotimah. "Attention guided sp...	<1%
119	Publication	Amanda Ramcharan, Tomislav Hengl, Dylan Beaudette, Skye Wills. "A Soil Bulk De...	<1%
120	Publication	Florin Nenciu, Mihai Gabriel Matache, Iuliana Gageanu, Ioan Catalin Persu, Florin ...	<1%
121	Submitted works	Galgotias University, Greater Noida on 2026-06-14	<1%
122	Publication	Jossy George, Kamal Upreti, Ramesh Chandra Poonia, Ankit Gautam, Danish Nad...	<1%

123	Publication	Patience Chizoba Mba, Opegbemi Matthias Busoye, John Temitope Ajayi, Judith N...	<1%
124	Publication	Sundaraj Kannan Susee, Shenbagaramasubramanian Shenbaga Vadivu, Muruges...	<1%
125	Submitted works	University of Sheffield on 2010-05-09	<1%
126	Submitted works	University of Sheffield on 2026-03-17	<1%
127	Submitted works	University of Wales Institute, Cardiff on 2026-08-01	<1%



Analyzing Soil Health Indicators for Agricultural Sustainability Using Machine Learning

Prepared By

Name: Alaa Mohammed Jawish

ID No.: 221002192

Name: Khadija Ehab Sadeq

ID No.: 221002423

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Mohamed Eassa

College of Informatics, Midocean University, Comoros

Email:

221002192@st.midocean.edu.km

221002423@st.midocean.edu.km

Acknowledgements

We would like to thank our supervisor, Dr. Mohamed Essa, for his continuous guidance and valuable feedback throughout this research. His expertise in both artificial intelligence and Soil health studies has greatly influenced the direction of this work. His comments challenged us to think more critically and approach the research from a scientific perspective rather than a purely technical one.

We are also thankful to the faculty members at the College of Informatics, Midocean University, for providing a supportive academic environment that inspired independent thinking and research development.

Finally, we would like to express our appreciation to our families for their patience, encouragement, and support during this journey.

Abstract

Due to the rising human population, in correlation the need for food increases making new issues emerge because not only for climate change but also for depletion of natural resources. Agricultural sustainability is now a worldwide problem, and in order to sustain agricultural productivity, soil health is crucial because it promotes crop growth, water retention, nutrient availability and ecosystem stability. Traditional techniques for evaluating soil health are time-consuming, frequently expensive and challenging to consistently implement over broad agricultural regions. Additionally, agricultural systems produce a multiplying volume of data that cannot be effectively analyzed using traditional methods as technologies continue to advance. Because of this, machine learning surfaced as one of the most useful tools humans use to evaluate complicated soil data and help with more trustworthy agriculture decision making. This study aims to analyze multiple soil health indicators by machine learning methods in order to identify the most influential indicators related to agriculture sustainability and provide a more efficient and reliable way to have sustainable soil management.



Content

	Acknowledgements	1
	Abstract	2
18	List of Abbreviations	4
	List of Figures	6
	List of Tables	7
	Chapter 1: Introduction	8
	1.1 Overview	8
	1.2 Problem Definition	9
	1.3 Thesis Objectives	10
	1.4 Thesis Contribution	10
32	1.5 Thesis Organization	10
	Chapter 2: Background and Related Work	11
	2.1 Introduction	11
	2.2 Background	11
	2.2.1 Soil Health Indicators	12
	2.2.2 Physical Indicators	13
	2.2.3 Chemical Indicators	13
	2.2.4 Biological Indicators	14
	2.2.5 Machine Learning Algorithms (ML)	14
	2.2.6 Evaluation Metrics	22
	2.3 Related work	24
16	2.3.1 Classification-Based Approaches for soil health assessment	24
16	2.3.2 Regression-Based Approaches for soil health assessment	26
	2.3.3 Related work summery	30
	References	35

List of Abbreviations

Abbreviation	Full term
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DNN	Deep Neural Network
RF	Random Forest
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
DT	Decision Tree
LR	Logistic Regression
XAI	Explainable Artificial Intelligence
SOC	Soil Organic Carbon
SOM	Soil Organic Matter
CEC	Cation Exchange Capacity
EC	Electrical Conductivity
pH	Potential of Hydrogen
N	Nitrogen
P	Phosphorus
K	Potassium
PTF	Pedotransfer Function
RMSE	Root Mean Square Error
MAE	Mean Absolute Error

R²	Coefficient of Determination
RFR	Random Forest Regression
RFC	Random Forest Classification
SVR	Support Vector Regression
PLSR	Partial Least Squares Regression
MLP	Multi-Layer Perceptron
SSL	Self-Supervised Learning
NIR	Near Infrared
MIR	Mid Infrared
GIS	Geographic Information System

List of Figures

Figure 2.1: Random Forest Architecture

Figure 2.2: Support Vector Machine Optimal Hyperplane

Figure 2.3: The Structure of Decision Tree

Figure 2.4: Artificial Neural Network Layers

Figure 2.5: Architecture of Deep Neural Network with multiple hidden layers

Figure 2.6: Extreme Learning Machine Structure

Figure 2.7: Gradient Boosting Machine Workflow

Figure 2.8: XGBoost Tree Structure

86

List of Tables

Table 2.1: The **Related Work Summary** on Machine Learning Applications for Soil Health Assessment

Chapter 1: Introduction

1.1 Overview

The need for agricultural sustainability has grown to a significant issue globally as food demand has risen and climate change and natural resource depletion have become a concern. Soil health is considered as one of the most crucial aspects to ensure the productivity of agriculture, as it affects many processes including nutrient availability, crop growth, water retention and ecosystem stability. Therefore, over time, soil health has become a crucial component of sustainable agriculture. With growing demands for sustainable agriculture, the need for ongoing soil monitoring and soil management has become even more critical over the past few years [1].

Soil health meaning describes the soil to support plants, animals and humans, and function in an environmentally healthy way as a living biota [2]. Commonly used measures of soil health include: soil moisture, pH, organic carbon, nutrients, electrical conductivity, microbial activity and soil structure [1]. These indicators give information of the state of soils and their capacity to support agriculture [2].

Soil health assessment is often done traditionally using field observations and laboratory testing methods. These have the advantage of being accurate but they are time consuming, expensive and hard to apply to large agricultural areas on a continuous basis [1]. In addition, sensor use, remote sensing and precision agriculture systems have led to tremendous amounts of agricultural data that need to be analyzed and interpreted effectively. Such data can be very difficult to analyses using traditional approaches and can delay timely agricultural decisions [3].

Effective methods for enhancing soil analysis and agricultural management have been made possible by the combination of artificial intelligence (AI) and machine learning (ML) [3]. Large-scale agricultural data analysis employing machine learning approaches reveals complex relationships between soil indicators and produces accurate forecasts about soil conditions [1]. Models like Random Forest, Neural Networks, and Support Vector Machines have shown a bright future in soil data analysis and agricultural decision-making [3].

Although soil science has made significant strides towards using machine learning to benefit agriculture, much of the existing literature is limited to specific soil characteristics, including soil fertility, soil moisture, or soil texture [1]. There is relatively little focus on the holistic

examination of many soil health indicators and how they all contribute to agricultural sustainability [1][3]. Furthermore, the relative weightage of various soil indicators is difficult to be determined because of the complexity of soils and the interaction between the different biotic, chemical and physical **properties of soil** [2].

By **using machine learning** analysis on soil health indicators, this study seeks **to identify the** most important **indicators**, investigate **the** correlations **between the** various **indicators**, and support sustainable agriculture management methods.

1.2 Problem Definition

Soil health evaluation is critical for sustainable farming in order to keep the productivity of agriculture [2]. The traditional way of analyzing the health of soils has been to use laboratory analysis and field measurements of many biological, physical, and chemical measurements. While important, these methods also are time-consuming, expensive, and hard to apply on a continuing basis over large agricultural areas [1].

Apart from soil health moisture, pH, nutrient availability, organic matter, and microbial activity, several indicator parameters also affect the soil health factors [2]. All of these interactions are complex, which makes it difficult to accurately monitor soil conditions using traditional analytical methods [1]. With the growing development of precision agriculture, sensor technologies and remote sensing systems, there is an increasing volume of soil data with the need for more efficient and intelligent analysis [1][3].

While a number of research **on the use of machine learning in agriculture** have been conducted, **the** majority **of** these studies have focused on a single soil characteristic rather than the concurrent assessment of many soil health indices. In addition, it is difficult to identify what are the most influential indicators that can contribute to agricultural sustainability [1][3][4].

From the discussion above, it is obvious that the main challenge in this thesis is the absence of an integrated machine learning solution that can analyze multiple soil health indicators, determine their relative importance and enable the implementation of sustainable agriculture decisions.

1.3 Thesis Objectives

This research aims to analyze soil health indicators using machine learning techniques and support agricultural sustainability through improved soil management practices.

o The objectives of this study are:

- To identify and investigate the key soil health indicators associated with agricultural sustainability.
- To collect, preprocess, and analyze soil data obtained from available datasets.
- To apply machine learning techniques to analyze the relationships among soil health indicators.
- To identify the most influential soil health indicators affecting agricultural sustainability.
- To evaluate and compare the performance of different machine learning models for soil health analysis.
- To provide data-driven insights that support sustainable agricultural decision-making and soil management practices.

1.4 Thesis Contribution

This thesis is expected to contribute to the fields of sustainable agriculture and artificial intelligence through the following:

- Developing a machine learning-based framework for analyzing multiple soil health indicators.
- Identifying the relative importance of different soil health indicators and their impact on agricultural sustainability.
- Providing a comparative evaluation of machine learning techniques used for soil health analysis.
- Supporting farmers and agricultural decision-makers through data-driven recommendations for sustainable soil management.
- Contributing to the growing application of artificial intelligence in precision agriculture and environmental sustainability.

1.5 Thesis Organization

The rest of this thesis is organized as follows.

Chapter 2 discusses the theoretical background and relevant research areas.

Chapter 3 presents the proposed methodology.

Chapter 4 analyzes the experimental results.

Chapter 5 concludes the study and suggests directions for future work.

Chapter 2: Background and Related Work

2.1 Introduction

This section presents a brief introduction and background information of previous studies about Soil health assessment and machine learning applications in agriculture sustainability. Some of the key ideas of soil health, soil health indicators, and agricultural sustainability will be demonstrated in this chapter. In addition to the machine learning methods that have been used for soil analysis and agriculture decision-making.

Effective Soil fertility management plays a critical role in achieving sustainable agricultural production. It aids farmers in preserving the soil's capacity to supply plants with the right quantity of nutrients when needed. Recently, soil analysis has been done using machine learning approaches. These techniques can process large amounts of agricultural data and help in making decisions related to soil and crop management [6].

However, many existing studies focus on some specific soil properties such as soil fertility, soil moisture, or nutrient prediction. In addition, different machine algorithms have been used with different datasets and evaluation methods, making it difficult to find the best approaches for soil health assessment.

Therefore, the background ideas of soil health, agricultural sustainability, and machine learning methods utilized in agriculture are covered in this chapter. In addition, it reviews previous work in detail in terms of methodologies, datasets used, and evaluation techniques in order to identify their strengths, limitations, and potential research gaps related to soil health analysis.

2.2 Background

Soil health has a significant role in plant growth, maintains environmental health quality, and improves sustainable agriculture production. Healthy soil improves crop productivity, nutrient cycling, water availability, and ecosystem stability. Overall, soil health is important not just for crops and food production but also for protecting the environment and assisting in climate change.

The functions of healthy soil are critical for environment protection and sustaining agriculture systems. It gives the plants essential nutrients, improves water filtration and storage, promotes biological activity, and maintains a stable soil structure [2].

Numerous biological, physical, and chemical factors affect the health of the soil. physical elements such soil structure, texture, and water-holding capacity. The chemical factors such as

potential pollutants, soil pH, organic carbon, nutrient availability, and electrical conductivity. Biological factors are also essential in maintaining soil functions. Examples of biological properties include microbial biomass and microbial activity. Based on this there is no single indicator that can fully describe soil health. For this reason, soil health assessment usually depends on evaluating a group of different indicators to give a more reliable assessment [2].

The biological, chemical, and physical state of the soil are among the indicators used to assess soil health. These indicators provide information on the soil's capacity to sustain plant growth and ecosystem services. Different indicators are combined to provide accurate assessment of soil condition [2].

Sustainable agriculture aims to provide enough food as demands continue to grow while protecting natural resources. It focuses on improving agricultural productivity through the efficient use of soil and other resources without affecting the needs of people in the future. Sustainable agriculture now has progressively adopted advanced technologies. To enhance farm management and facilitate well-informed decision-making, these technologies include remote sensing, the internet of things (IOT), and precision agriculture [3].

The management of agricultural systems has been changed because of the rapid development of digital technologies. Modern farming generates large amounts of data from drones, satellites, sensors, and other smart devices. Traditional methods are no longer sufficient for analyzing these data. Thus, ML has integrated as an effective method for processing agriculture data, identifying hidden patterns, and supporting accurate and timely decision making in sustainable agriculture [3].

Researchers have used different machine learning algorithms in agriculture to study complex data and make predictions more accurate. These algorithms examples are Decision Tree, Support Vector Machine, Extreme Gradient Boosting (XGBoost), multilayer perceptron (MLP) neural networks, Artificial Neural Networks (ANN), and Random Forest. These algorithms have been applied in various agricultural practices, such as soil analysis, disease detection, crop yield prediction, and precision farming. These algorithms are useful tools for assisting with sustainable agriculture management since they can examine big information and find intricate linkages. [3]. Additionally, smart agriculture is being revolutionized by the integration of IoT and ML. Researchers have recently looked into using these technologies to automate agricultural tasks like pest control, irrigation management, weather monitoring, and soil condition monitoring. [3].

2.2.1 Soil Health Indicators

Measurable soil characteristics are known as soil indicators. The indicators are used to assess the condition of the soil and determine how effectively it sustains plant growth and

environmental quality. These indicators provide valuable information about the biological, chemical, and physical characteristics of the soil. To assess soil condition, a combination of different indicators is used. [2]

2.2.2 Physical Indicators

It describes the physical condition of the soil such as soil moisture, soil texture, soil structure, and bulk density. These indicators effect water movement, root development, soil aeration, and resistance to erosion. [2]

1- Soil Moisture. It refers to how much water is stored in the soil. Its directly affects seed germination, nutrient transport, root development, and crop productivity.[2]

2- Soil Texture. Soil texture describes how soil particles are organized and how large the space between them. [1], [2]

3- Bulk Density. It is defined as what is the mass of the dry soil per unit volume which used to evaluate soil compaction, which impact air movement within the soil, root penetration, and water infiltration.[2]

4- Soil Structure. It describes how soil particles are arranged into aggregates to enhance root penetration, water infiltration, aeration, and resistance to erosion.[2]

2.2.3 Chemical Indicators

Chemical indicators describe the chemical composition of soil and provide information about its fertility and nutrient availability. These indicators including soil organic carbon, soil pH, potassium, nitrogen, phosphorus, and electrical conductivity (EC) impact both microbial activity and plant growth.[1][2]

1- Soil pH. It is an important indicator that describes the acidic or alkaline of soil. Extreme pH values can reduce nutrient uptake and negatively influence crop productivity. [2]

2- Soil Organic Carbon (SOC). It represents the amount of organic carbon that stored in the soil it is essential in the global carbon cycle and climate change.[2]

3- Nitrogen (N). Nitrogen is essential for both photosynthesis and the production of proteins. Insufficient nitrogen levels can reduce crop yield.[2]

4- Phosphorus (P). Phosphorus is an essential nutrient that supports root development, energy transfer, and flowering.[2]

5- Potassium (K). It is a crucial nutrient that controls a number of plant functions, including as water balance, enzyme activity, and stress tolerance.[2]

6- Electrical Conductivity (EC). It is used to evaluate soil salinity, which can have an impact on the amount of water that plants absorb, by measuring the concentration of soluble salts in the soil.[2]

2.2.4 Biological Indicators

It evaluates the living components and provides information about microbial activity. Common biological indicators include soil respiration, soil organic matter, and microbial activity. [1], [2].

1- Soil Organic Matter (SOM). It consists of decomposed animal and plant residues found in the soil and it improves soil structure.

2- Microbial Activity. Microbial activity refers to the activity of microorganisms like fungi and bacteria in the soil high microbial activity mean healthy soil.[2]

3- Soil Respiration. The amount of carbon dioxide released by soil microorganisms during the breakdown of organic materials is measured by soil respiration.

2.2.5 Machine Learning Algorithms (ML)

The management of agricultural systems has been changed because of the rapid development of digital technologies. Modern farming generates large volumes of data obtained by satellites, drones, sensors, and other smart devices. Traditional methods are no longer sufficient for analyzing these data. Thus, machine learning has become an effective method for processing agriculture data, identifying hidden patterns, and supporting accurate and timely decision making in sustainable agriculture [3].

Numerous machine learning (ML) techniques have been applied in agricultural research to assess challenging data and increase prediction accuracy. Examples of these methods include Artificial Neural Networks (ANN), Decision Trees, Support Vector Machines, Random Forests, Multilayer Perceptron (MLP) neural networks, and Extreme Gradient Boosting (XGBoost).

Several agricultural applications, including soil analysis, crop yield prediction, precision farming, and disease detection, have made use of these algorithms. They can be helpful tools for supporting sustainable agriculture management that analyze large volumes of data and uncover complex relationships [3]. In addition, Smart agriculture is being revolutionized by the combination of IoT and ML. In more recent times, scientists have looked into using these technologies to automate other farm tasks like soil condition monitoring, weather monitoring, irrigation management, and insect control [3].

1. Random Forest (RF)

An ensemble algorithm is an aggregation of multiple decision trees with the aim of boosting the accuracy of data classification and prediction. This approach is particularly reliable as it can handle noisy data, is not prone to overfitting and is good for large data and input variables. When using an ensemble algorithm, the following formula can be used to calculate the regression results:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B f_b(x) \quad (1)$$

where B is the total number of trees and $f_b(x)$ represents the estimation of the b -th decision tree.

Random Forest

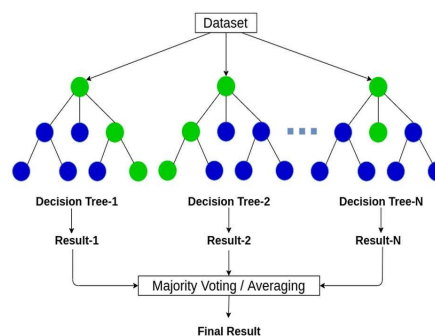


Figure 2.1: Random Forest architecture

2. Support Vector Machine (SVM)

is a type of supervised ML algorithm that is applied for regression and classification analysis. It is based on determining the best separating hyperplane that divides the space of classes SVM uses kernel functions for the best approximation of the surface between the categories. In the linear model case.

the following expression gives the function:

$$\min_{w,b,\xi} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \right) \quad (2)$$

subject to the constraints $y_i(w \cdot x_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ where w represents the weight vector c is the regularization parameter and ξ_i represents the slack variables for margin violations.

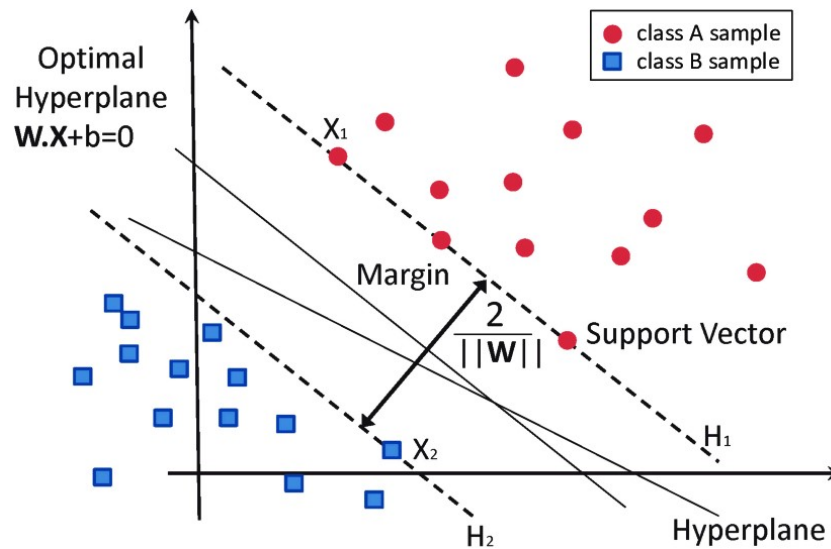


Figure 2.2: Support Vector Machine optimal hyperplane

3. Decision Tree (DT).

Is considered as a supervised learning method which models the data as a decision tree based on a few decision rules. This algorithm is quite simple to interpret and it needs little data preparation and is frequently applied for the soil classification and prediction problems. For the classification tasks. The decision trees pick the best splitting criteria. according to the lowest Gini Impurity.

which is calculated by the equation below:

$$Gini = 1 - \sum_{i=1}^c (p_i)^2 \quad (3)$$

Where p_i is the probability of an item belonging to class i , c = the total no. of classes.

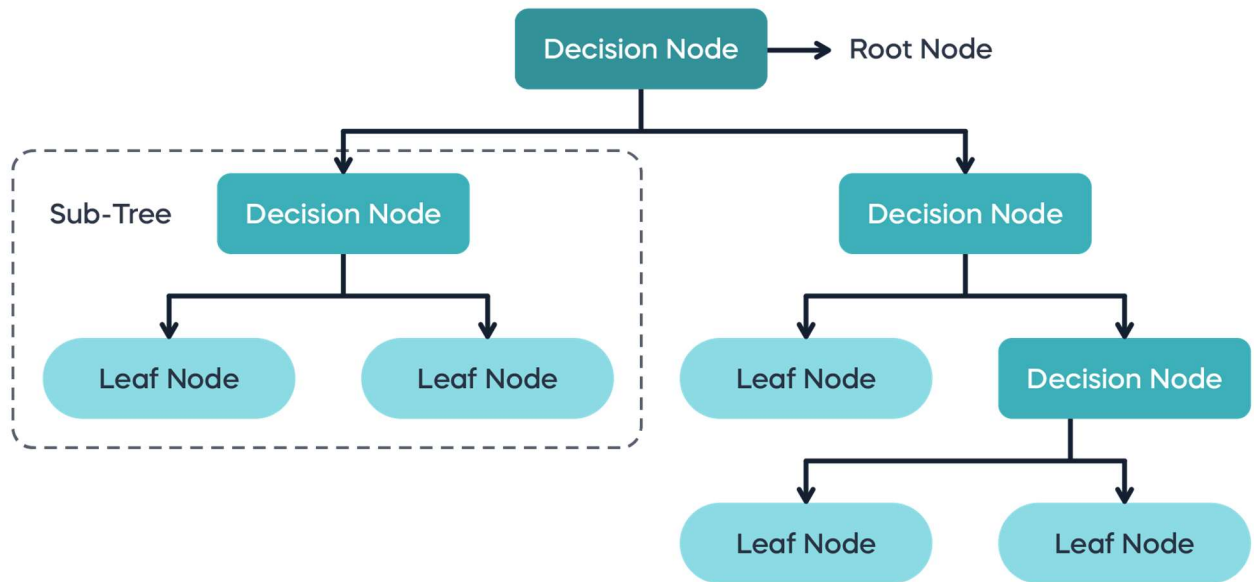


Figure 2.3: The structure of Decision Tree

4. Artificial Neural Network (ANN).

This artificial neural network consists of the input layer, hidden layer and output layer.

Artificial neural networks (ANNs) can well capture the intricate nonlinear interactions of various chemical and physical characteristics of soil health indicators.

A single neuron's output is given by the following formula:

$$y = f(\sum_{i=1}^n w_i x_i + b) \quad (4)$$

where b is the bias term, x_i represents the input features w_i denotes the corresponding weights and f is the activation function (such as ReLU or Sigmoid).

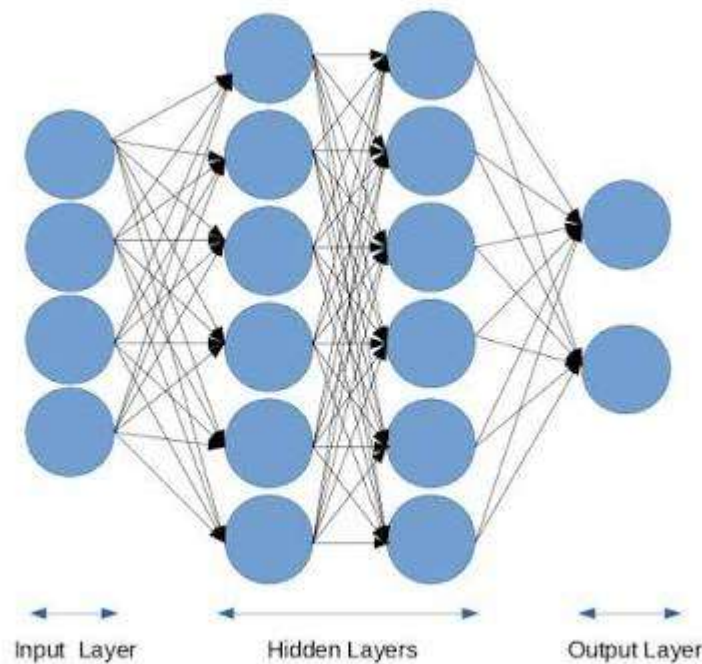


Figure 2.4: Structure of Artificial Neural Network with input, hidden, and output layers

5. Deep Neural Network (DNN).

is an advanced form of ANN having many layers, enabling it to learn multiple levels of representation from data. The network involves more than one hidden layer in order to learn nonlinear combinations of the input data better thus allowing the algorithm to automatically learn intricate functions from high-dimensional agricultural data.

The output of any given layer l in the network can be defined by the equation below:

$$h^l = \sigma(W^{(l)}h^{(l-1)} + b^{(l)}) \quad (5)$$

where h^l represents the activations of layer l and $W^{(l)}$ is the weight matrix for layer l and $b^{(l)}$ is the bias vector and σ is the non-linear activation function.

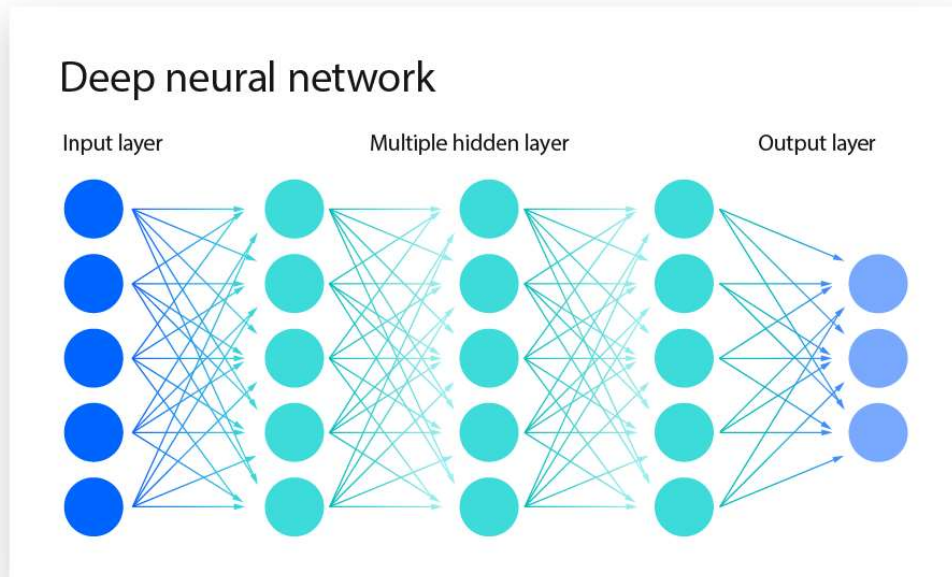


Figure 2.5: architecture of Deep Neural Network with multiple hidden layers

6. Extreme Learning Machine (ELM).

The Extreme Learning Machine (ELM) is a type of feed-forward neural network where the weights of the hidden layers are fixed and random, and the weights of the output layer are computed analytically. since there is no need for iterative training. The calculation becomes extremely fast. Compared to classical algorithms ELM can be several times faster which makes it suitable for large sets of data such as the soil data set discussed above.

The formula that is used for calculation of weights is expressed below:

$$\beta = H^{\dagger} Y \quad (6)$$

where H is the hidden layer output matrix $H^{\dagger}$ its generalized inverse and Y the target matrix.

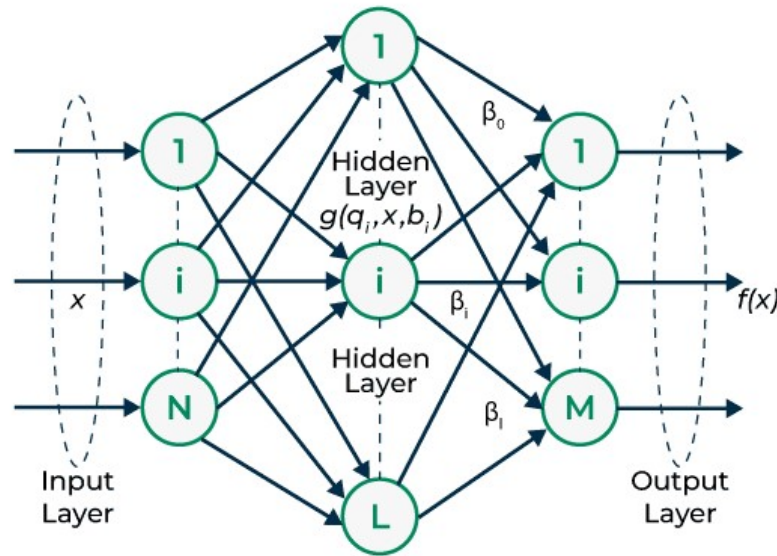


Figure 2.6: Extreme Learning Machine structure

7. Gradient Boosting Machine (GBM).

It builds trees sequentially in which every subsequent tree is trained to fix the mistakes of the one before it. As opposed to the random forest algorithm, GB sequentially enhances the prediction accuracy by concentrating on minimizing residuals from the trees and therefore can capture complex relationships that exist between soil indicators.

The model is updated by decreasing the loss function:

$$f_m(x) = f_{m-1}(x) + \arg \min_h \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + h(x_i)) \quad (7)$$

where $f_m(x)$ is the model at iteration m and $h(x)$ is the weak learner fitted to the residuals.

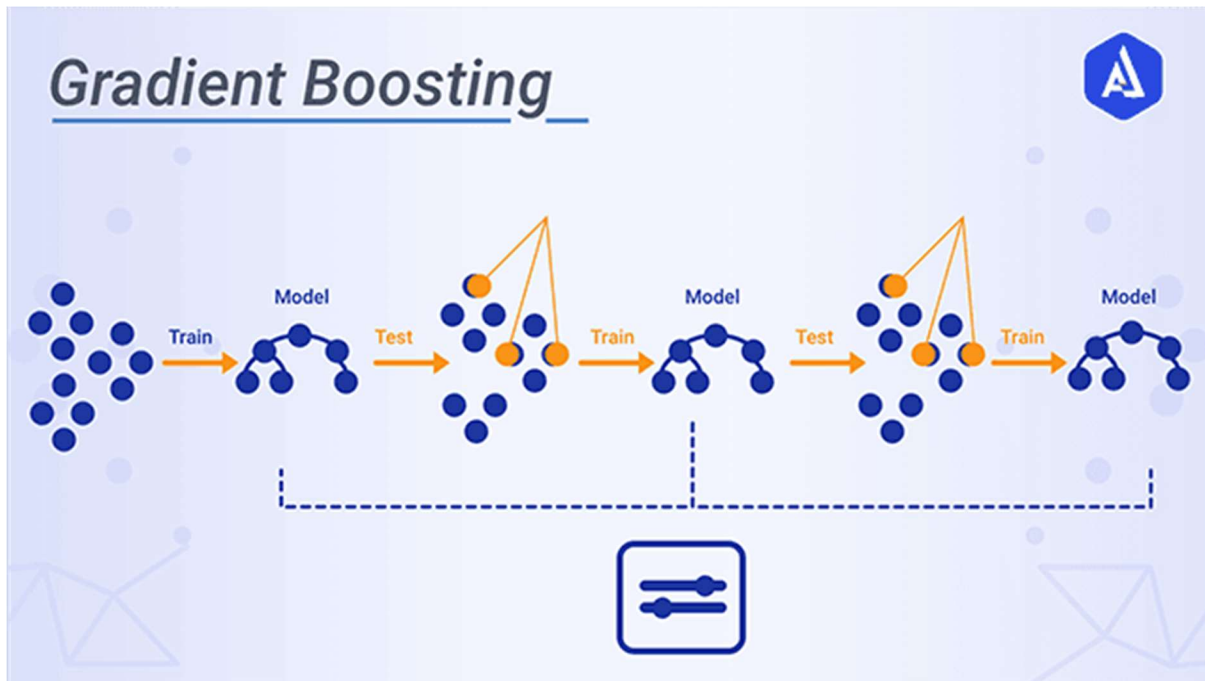


Figure 2.7: Gradient Boosting Machine workflow

8. XGBoost.

It is an optimized regularized application of gradient boosting which is primarily designed to be highly efficient and accurate. It includes built-in regularization to prevent model complexity and overfitting and automatically manage missing data. An important feature for many real-life applications such as agricultural yield prediction which is the main goal of this research.

The objective function minimized at every step is:

$$Obj^{(t)} = \sum_{i=1}^n L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (8)$$

where L = the loss function f_t the new tree and $\Omega(f_t)$ penalizes tree complexity.

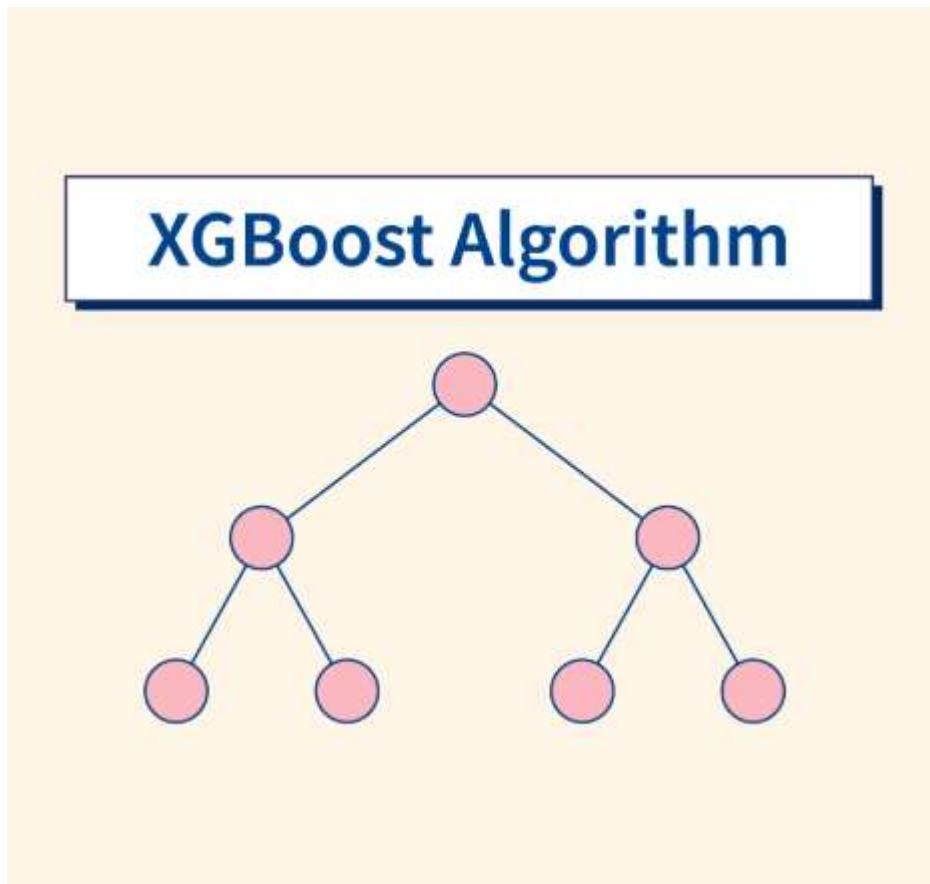


Figure 2.8: XGBoost tree structure

9. Convolutional Neural Network (CNN)

It is a deep learning model used to structured data processing as a grid, be it images or, in this case, soil data. They can be successfully applied to data mining problems such as the one with the hyperspectral soil readings since they allow to automatically learn important features or filters and without performing an explicit search.

The convolution operation works as the sliding of the filter above the data matrix:

$$S(i, j) = (I * K)(i, j) = \sum_{m, n} I(i - m, j - n) K(m, n) \quad (9)$$

where $S(i, j)$ is the resulting output feature map.

2.2.6 Evaluation Metrics

To accurately evaluate the performance and prediction capabilities of the implemented ML models, specific evaluation metrics will be utilized. These mathematical metrics provide a better understanding of the error margins and the precision of the model predictions:

- Coefficient of Determination (R^2):** Evaluates how good the model predicts the variance of the target variable. A higher R^2 indicates a better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

Root Mean Square Error (RMSE): determines the square root of the average of the squared differences between the actual values and predicted. It is highly sensitive to huge errors.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (11)$$

Mean Absolute Error (MAE): determines the average of the absolute differences between the actual values and the predicted values, offering a linear evaluate of the average error.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (12)$$

Accuracy: this metric refers to the ratio of the number of correct classifications made by the algorithm and the total number of classifications:

$$Accuracy = \frac{TP+T}{TP+TN+FP+F} \quad (13)$$

Precision: defines the ratio of samples predicted as a specific class correctly predicted by the algorithm. if the precision is high, it is less likely that the soil health is predicted healthy when it is not:

$$Precision = \frac{TP}{TP+FP} \quad (14)$$

Recall: is the ratio of samples of a specific class predicted correctly by the algorithm. In other words. it indicates whether the algorithm predicts healthy soils as healthy and unhealthy soils as unhealthy:

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

F1-Score: represents the harmonic mean of Precision and Recall and is used when the distribution of classes is not equal:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (16)$$

2.3 Related work

Several studies have applied machine learning techniques in agriculture and soil health analysis. These studies focused on different soil properties. While the results were promising, most of these studies focused on specific soil parameters rather than analyzing multiple soil health indicators together. This makes it important to review these studies to understand the existing research gap and current research trends.

2.3.1 Classification-Based Approaches for soil health assessment

Rao and Reddy (2026) proposed the AgriMind Intelligent Fertility Predictor (AMIFP) for classification of the soil fertility using deep learning techniques. The proposed framework integrated a novel preprocessing method (SIC), an attention-based deep learning model (ALADN), and a genetic hyperparameter optimization algorithm (GPSHO). The study used 880 soil samples from the Kaggle dataset, The dataset was separated into training, validation, and testing categories (70,10,20). According to the findings, the model performed well in classifying soil fertility, achieving accuracy of 98.65%, precision of 98.73%, recall of 98.65%, and F1-score of 98.70%. However, the evaluation was limited to a single public dataset which may affect in generalize the model to different soil conditions.[7]

Behera et al. (2026) compared a few classifiers of machine learning to make the prediction of soil fertility effective. In this study, sixteen machine learning classifiers, such as ensemble-based, instance-based, learner-based, and probabilistic-based classifiers were compared to estimate the soil fertility. The study was based on 880 soil samples available from Kaggle with 12 physicochemical soil properties. The result indicated that the RF model gave best accuracy with the rate of 90.91% over all other models. The performance of the model was however found to be relatively low for one of the minority classes, suggesting that the model faced problems with class imbalance and overlapping feature distributions. [8]

Kullayamma and Venkatanaresh (2024) suggested a deep learning model for automatically recommending crops depending on climate, fertilizer, and soil. Real-time Maharashtra datasets and Kaggle datasets were used in the study. The Concurrent Excited Gated Recurrent Unit (CEGRU) model was suggested by the study. The Hunter-Prey Optimization (HPO) algorithm was used for parameter optimization in order to enhance the model's performance. The proposed approach was evaluated by using many performance metrics, including positive

predictive value (PPV), accuracy, kappa, time complexity, RMSE, recall, and sensitivity. The result showed that the framework achieved high classification results with an accuracy value of 99.2% and 99.3% for Maharashtra and Kaggle dataset. However, the study focused on recommending suitable crops according to the climatic conditions and soil fertilizer, while the proposed research focuses on analyzing multiple soil health indicators.[9]

Badshah et al. (2024) proposed ML models for agricultural prediction tasks using both classification and regression frameworks based on two different datasets. In the initial crop suggestion, the authors utilized a Kaggle dataset that included soil and ambient characteristics like temperature, humidity, pH, and nutrient levels. Decision Tree (DT), Extra Tree Classifier (ETC), Random Forest (RF), K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), and Gaussian Naive Bayes (GNB) were among the classification techniques that were compared. With an accuracy of 99.7%, the Random Forest (RF) model uses Explainable AI (XAI), feature engineering, and K-fold cross-validation to get the best classification results. Additionally, the study used World Bank and Food and Agriculture Organization (FAO) records to apply regression models for wheat yield prediction. The outcome demonstrated that Support Vector Regressor (SVR) with hyperparameter optimization and K-fold cross-validation performed best, achieving an accuracy of 99.9% with R2 Score. Additionally, model predictions were interpreted using Explainable AI techniques like Feature Importance and LIME. However, the study focused mainly on crop recommendation and yield prediction, where soil parameters were used as input features, while the proposed research focuses on analyzing multiple soil health indicators and identifying the most influential factors affecting agricultural sustainability using machine learning techniques.[10]

Sarangi et al. (2024) created a machine learning-based method for evaluating soil fertility utilizing topography data, temperature, rainfall, moisture, nutrient level, nitrogen, phosphorus, potassium, and phosphorus. The study utilized a publicly available Kaggle dataset. The study's goal was to categorize soil as either fertile or non-fertile. The best classifier is determined using four classification algorithms: Logistic Regression (LR), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Decision Tree (DT). The results indicated that the Decision Tree classifier achieved the highest accuracy of 89%, while Logistic Regression and KNN achieved approximately 90% performance in some evaluation metrics. However, the study focused on binary soil fertility classification and did not analyze multiple soil health indicators or identify the most influential factors affecting agricultural sustainability. [11]

Chandra et al. (2023) discussed utilizing Random Forest classification to create an explainable AI (XAI) model for predicting soil fertility. It used several soil characteristics including nitrogen, pH, potassium, phosphorus, electrical conductivity, and organic carbon. The LUCAS 2018 Topsoil dataset was utilized in this investigation. To explain the outcome, the authors integrated Random Forest, SHAP, and k-means clustering. The outcome demonstrated the

efficacy of explainable machine learning in soil fertility measurement, with an accuracy of 97.02%. However, this study focused on soil fertility prediction only, and did not analyze multiple soil health indicators [12].

2.3.2 Regression-Based Approaches for soil health assessment

Ghahremani et al. (2026) devised a high-resolution CONUS soil inorganic carbon (SIC) map technique using machine learning models trained with climate and land surface factors as well as the ISRIC World Soil Information Service (WoSIS) database at 30 m resolution. Two machine learning models, Random Forest Regression (RFR) and Random Forest Classification (RFC) were used to estimate the SIC distribution. The prediction performance of the Random Forest Regression model was better than the Classification model with an RMSE of 15 kg/m² and accuracy of 0.56. Soil pH was found to be an important factor affecting soil inorganic carbon distribution in the study. The study did focus on mapping soil inorganic carbon, but the proposed research will aim to examine a range of soil health indicators to gain a clearer understanding of their role in ensuring sustainable farming.

Sun et al. (2026) suggested a framework for self-supervised multi-fidelity learning that uses spectral data to predict various soil parameters. The study used the dataset from the (USDA-NRCS-NSSC-KSSL) database, a paired NIR–MIR dataset, and the North American Proficiency Testing (NAPT) dataset. The proposed framework combined Self-Supervised Learning (SSL) with a Multi-Layer Perceptron (MLP) and compared its performance with conventional MLP and Partial Least Squares Regression (PLSR) models. The MIR-SSL-MLP model achieved best prediction performance for several soil properties, obtaining R² of 0.95 for total carbon, 0.91 for total nitrogen, 0.93 for estimated organic carbon, and 0.79 for pH. However, the proposed framework relied mainly on soil spectral datasets and did not integrate environmental or remote sensing variables for soil health assessment. [14]

Khatun et al. (2026) investigated the potential of spectroscopy combined with machine learning as a non-destructive and effective way to forecast soil organic carbon (SOC) using mid-infrared spectroscopy (MIR) data in the 600–4000 cm^{−1} spectral region. The study utilized the Open Soil Spectral Library (OSSL), which integrates soil spectral data from several international databases, including KSSL, LUCAS, RaCA, AfSIS, and ICRAF–ISRIC. Mid-infrared (MIR) spectral data was used to create Random Forest (RF) and Partial Least Squares Regression (PLSR) models, and Shapley Additive Explanations (SHAP) was used to understand the contribution of individual spectral wavelengths. With R² of 0.98 and RMSEP (a root mean square error) of 1.45%, the RF model had the best performance following Grid-Search hyperparameter optimization. However, the study focused on predicting soil organic carbon from spectral information while the proposed search focuses on using Machine Learning

Algorithms to analyze multiple soil health indicators and identifying the most effective factors in agricultural sustainability.[15]

Aouabed et al. (2026) examined the potential of machine learning techniques to assess the functional and taxonomic composition of soil microbial communities by examining data from two publicly accessible soil microbiome datasets: (1) Information gathered by Averill et al. and (2) Information taken from the NEON database, which assessed a number of deep learning and machine learning models. A number of machine learning models were assessed and contrasted, including Support Vector Regression (SVR), Gradient Boosting (GB), Multilayer Perceptron (MLP), Random Forest (RF), Ridge Regression, Lasso Regression, and k-Nearest Neighbors (k-NN). According to the experimental findings, ML models outperformed deep learning models on the majority of prediction tasks, especially when training data was scarce. For the bacterial dataset, Gradient boosting had the greatest prediction performance with an R2 value of 0.57; for the fungal functional groups, Random Forest and Gradient boosting models produced an R2 value of 0.45. However, the study focused on predicting soil microbiome composition as a biological indicator of soil health, whereas the proposed research analyzes multiple physical and chemical soil health indicators to identify the most influential factors affecting agricultural sustainability.[16]

Kassai et al. (2025) created pedotransfer functions (PTFs) to calculate the distribution of particle sizes based on soil characteristics. The Hungarian Detailed Soil Hydrophysical Database, which has 3,970 soil profiles, was used to train and evaluate the prediction models using the Random Forest approach. The study also made use of the publicly accessible NATASA database, which has 1,372 soil profiles. The Random Forest Kriging (RFK) was used to generate high-resolution digital soil maps for sand, silt, and clay fractions. The developed model showed that the R² values of the PTFs are 0.74 for sand content, 0.69 for clay, and 0.58 for silt. In addition, the RFK models accomplished R² values between 0.19 and 0.67 for clay, 0.49 to 0.62 for silt, and 0.69 to 0.74 for sand. However, the study focused on soil texture mapping rather than analyzing multiple soil health indicators affecting agricultural sustainability.[17]

Liu and Ma (2024) focused on predicting soil field capacity (FC) and permanent wilting point (PWP) based on parameters obtained from WoSIS database. The study predicted the hydraulic parameters of the soil using Gene-Expression Programming (GEP) models and Artificial Neural Networks (ANN). With a normalized root mean square error (NRMSE) of 19.9% for PWP prediction and 29.9% for FC prediction, the ANN model outperformed the GEP model, according to the model. Additionally, when compared to the traditional linear models, the ANN model decreased the mean absolute error (MAE) by 51.54% for the FC and 56.38% for the PWP. However, the study focused only on soil hydraulic properties and did not analyze multiple soil health indicators that affect agricultural sustainability.[18]

Nauman et al. (2024) developed predictive digital soil property maps for the conterminous United States using a Random Forest-based digital soil mapping framework with hybrid training sets. The study integrated publicly available datasets, including the National Soil Information System (NASIS), the National Cooperative Soil Survey (NCSS) Soil Characterization Database, and the Soil Survey Geographic Database (SSURGO), to generate 100 m resolution maps for several soil properties, including various, chemical parameters, physical parameters, carbon, textural fractions, and depth to restrictions. A hybrid training strategy used to increase training data. Model performance varied considerably across properties, depths, training data types, and validation approaches. The developed models achieved generally strong prediction performance, with R^2 values ranging between 0.5 and 0.7 under cross-validation. However, the study focused on digital soil mapping for the United States and did not investigate the relationships among multiple soil health indicators for agricultural sustainability.[19]

Zhai et al. (2024) focused on simultaneously predicting multiple soil properties by a partially shared multi-task learning model (PS-MTL-LUCAS). In order to enhance prediction accuracy and interpret significant spectrum information, the study suggested a multi-task deep learning architecture in conjunction with SHapley Additive exPlanations (SHAP). Seven soil parameters, including pH, phosphorus (P), nitrogen (N), potassium (K), organic carbon (OC), cation exchange capacity (CEC), and calcium carbonate (CaCO_3), were predicted using the LUCAS Topsoil Dataset (2009). The outcomes showed high prediction performance, achieving R^2 values of 0.952, 0.936, 0.945, 0.413, 0.837, 0.624, and 0.956 for CaCO_3 , N, pH, P, CEC, K, and OC respectively. However, the study was tested only on the LUCAS Topsoil Dataset using Vis-NIR spectroscopy data, and did not evaluate other soil datasets or under real field conditions.[20]

Saberioon et al. (2024) used deep learning methods to enhance soil organic carbon prediction. The total number of samples included in the study was 11441 from LUCAS Soil Database. A several models were compared: Fully Connected Neural Network (FCNN), One-Dimensional Convolutional Neural Network (1DCNN), Random Forest (RF) model and Stacked AutoEncoder (SAE). The SAE-1DCNN model accomplished the ideal prediction performance achieving an R^2 of 0.784 and an RMSE of 3.94. The study, however, only attempted to predict Soil Organic Carbon (SOC) and did not study multiple soil health indicators that impact agricultural sustainability.

Liu et al. (2023) used the Visible and Near Infrared Diffuse Reflectance Spectroscopy (Vis-NIR DRS) instrument to study the prediction of Soil Organic Carbon (SOC). The study compared various prediction techniques, such as Global Partial Least Squares Regression (PLSR), Local PLSR, k-means clustering with PLSR, and hierarchical clustering with PLSR, using the LUCAS EU-wide Topsoil Database. The outcome presented that the ideal

performance were provided by the local PLSR approach achieving R^2 value of 0.75, (RPIQ) of 2.846, and (RMSEP) of 13.38 g/kg however the model running time was about 30.05 s. same performance were provided by the k-means clustering technique achieving R^2 value of 0.75, RMSEP of 14.61 g/kg and RPIQ of 2.844, at only 4.52 s. However, the study focused only on Soil Organic Carbon prediction while the proposed research aims to analyze multiple soil health indicators to better understand their contribution to sustainable agriculture. [22]

Omondiagbe et al. (2023) developed an automated deep learning algorithm to predict soil texture. To improve model performance, the study employed Deep Convolutional Neural Networks (CNNs) with Adapted Population-Based Training (Adapted-PBT) and Bayesian Optimization. A dataset was used from the Kellogg Soil Survey Laboratory (KSSL) and the LUCAS Soil Library. With model efficiency coefficient (MEC) values of 0.90 for clay, 0.81 for sand, and 0.78 for silt, the results demonstrated that the PBT model offers the best prediction performance. However, the search based on predicting soil texture properties (sand, silt, and clay) and did not investigate other soil health indicators or their relationships.[23]

Piccoli et al. (2023) proposed an adaptive deep neural architecture for the prediction of multiple soil characteristics from the analysis of hyperspectral signatures. The study used the LUCAS soil dataset and a simulated PRISMA hyperspectral dataset to evaluate the model. The model is based on a flexible neural architecture that can backtrack to the spectral bands that most contribute to the prediction of a given variable and automatically adapt to the spectral library during analysis. The model accomplishes a total R^2 of around 0.75. More clearly, R^2 values exceeding 0.90 for pH and CaCO_3 , while organic carbon, nitrogen, clay, sand, silt, and cation exchange capacity achieved R^2 values between 0.70 and 0.89. However, the study relied mainly on hyperspectral information and did not include additional environmental or management factors.[24]

Rahmani et al. (2022) developed digital maps of soil organic matter (SOM) and cation exchange capacity (CEC) using LiDAR-derived terrain attributes and compare the prediction accuracy of SOM and CEC maps created by universal kriging, Cubist, and random forest with Soil Survey Geographic (SSURGO) database. The result showed comparable performance With R^2 values from 0.44 to 0.45 for SOM and from 0.39 to 0.44 for CEC. However, the study focused on mapping two soil properties within a local study area and did not investigate multiple soil health indicators or employ large-scale global soil datasets relevant to agricultural sustainability.[25]

Ramcharan et al. (2017) described a model to develop a soil bulk density pedotransfer function (PTF) using the Random Forest. The study used around 46,000 soil horizons collected from the publicly available National Cooperative Soil Survey (NCSS) Soil Characterization Database in the United States. The information contained hierarchical ecosystem land classifications and surficial materials. The outcome of a five-fold cross-validation presented

that the model (RMSPE) was (0.13 g cm⁻³) in average, while the (MPE) was (-0.001 g cm⁻³). In addition, the developed PTF was also applied to support soil organic carbon (SOC) stocks estimation by addressing missing bulk density information. However, the study focused on predicting soil bulk density only and did not analyze multiple soil health indicators or evaluate their relationships. [26]

2.3.3 Related work summery

Machine learning methods have been shown to be effective in studying various soil properties and in the decision-making process in agricultural practices. Most regression-based studies have been conducted to predict an individual soil property such as soil organic carbon, soil texture and hydraulic properties from spectral data or environmental variables. Most of these studies focused on a small number of soil indicators rather than on soil health in its entirety. These methods have had promising prediction performance, but most studies tested a small number of soil indicators, not soil health as a whole.

However, on the other hand, classification-based approaches have been used in soil fertility assessment and agricultural recommendation tasks. The machine learning classifiers were found to be capable of correctly categorizing soil conditions in these studies. However, there are many existing works that were limited to predefined classes (e.g., fertile/non-fertile classes) and did not investigate the contributions of each indicator of soil health.

Thus, there is a need for a holistic assessment approach that takes several soil health indicators simultaneously and establishes the most significant factors that are associated with sustainable agriculture.

Table 2.1: The Related Work Summary on Machine Learning Applications for Soil Health Assessment

<i>Ref</i>	<i>Author</i>	<i>Year</i>	<i>Technique Used</i>	<i>Accuracy</i>	<i>Dataset</i>	<i>Comparison with This Research</i>
[7]	Rao and Reddy	2026	SIC ALADN GPSHO	precision 98.73%, accuracy 98.65%, F1-score 98.70% recall 98.65%, and	Kaggle dataset containing 880 soil samples	using only one publicly available dataset
[8]	Behera et al.	2026	16 machine learning classifiers	RF (accuracy =90.91%	Kaggle dataset containing 880 soil samples	the model showed relatively low performance for one of the minority classes
[9]	Venkatanare sh and Kullayamma	2024	CEGRU HPO	accuracies of 99.2% and 99.3%	Kaggle + Maharashtra	focused only on recommending suitable crops based on soil, fertilizer
[10]	Badshah et al	2024	(ETC), (LR), (DT), (RF), (KNN), (GNB), and (SVM).	best performance by Random Forest by accuracy of 99.7% by use K-fold cross	crop recommendation dataset obtained from Kaggle, FAO	study focused mainly on crop recommendation and yield prediction
[11]	Sarangi et al.	2024	LR, KNN, NB, DT.	the DT best accuracy of 89%	Kaggle	focused on binary soil fertility classification
[12]	Chandra et al.	2023	k-means clustering Random Forest classifier, Shap	~97%	LUCAS 2018 topsoil dataset	Limited to fertility prediction
[13]	Ghahremani et al.	2026	RFR RFC	RFR with an RMSE of 15 kg/m ² , while RFC accuracy of 0.56.	ISRIC World Soil Information Service (WoSIS) database	Limited to mapping soil inorganic carbon
[14]	Sun et al.	2026	SSL, MLP compared performance with PLSR and conventional MLP models.	MIR-SSL-MLP model best performance R ² of 0.95 carbon, 0.91 nitrogen, 0.93 organic carbon, 0.79 for pH	(USDA-NRCS-NSSC-KSSL) database Testing (NAPT) dataset.	relied mainly on soil spectral datasets and did not integrate environmental or remote sensing variables for soil health assessment.

[15]	Khatun et al.	2026	PLSR RF SHAP	$R^2 = 0.98$ RMSEP = 1.45%	(OSSL) including KSSL, LUCAS, RaCA, AfSIS, and ICRAF-ISRIC	study only focused on predicting SOC from spectral information
[16]	Aouabed et al.	2026	RF KNN GB SVR Ridge Regression Lasso Regression Multilayer Perceptron (MLP) Regressor	The best overall $R^2=0.57$ by (GB)	Averill et al. Dataset NEON Dataset	focused on predicting soil microbiome composition
[17]	Kassai et al.	2025	Random Forest Random Forest Kriging (RFK)	PTF R^2 Clay = 0.69 Silt = 0.58 Sand = 0.74, RFK R^2 Clay 0.19–0.67 Silt 0.49–0.62 Sand 0.69–0.74	NATASA database	focused only on soil texture mapping rather than analyzing multiple soil health indicators
[18]	Liu and Ma	2024	(ANN) (GEP) LinearRegression for comparison	ANN model made the best for PWP with NRMSE (19.9%) while GEP with an NRMSE (29.9%)	WoSIS database.	focused only on soil hydraulic properties
[19]	Nauman et al.	2024	Random Forest	R^2 between 0.5 and 0.7 under cross-validation.	NCSS SSURGO / gSSURGO NASIS	focused only on digital soil mapping for the United States and did not investigate the relationships among multiple soil health indicators

[20]	Zhai et al.	2024	PS-MTL-LUCAS SHAP	achieving R^2 values of 0.945(Ph), 0.936(N), 0.413(P), 0.624(k), 0.837(CEC), 0.952(OC), and 0.956(CaCO_3)	LUCAS Topsoil Dataset (2009)	only on the LUCAS Topsoil Dataset using Vis-NIR spectroscopy data
[21]	Saberioon et al.	2024	1DCNN FCNN SAE RF	SAE-1DCNN made the best result with an R^2 of 0.784 and an RMSE of 3.94.	11441 samples collected from LUCAS Soil Database	Limited to predicting (SOC)
[22]	Liu et al.	2023	Global PLSR Local PLSR k-means clustering + PLSR Hierarchical clustering + PLSR	best results for local PLSR $R^2 = 0.75$, RPIQ) = 2.846 (RMSEP) = 13.38 g/kg	LUCAS EU-wide Topsoil Database	focused only on SOC prediction
[23]	Omondiagbe et al.	2023	(Adapted-PBT) Bayesian Optimization (CNN)	PBT best prediction with (MEC) of 0.90 for clay, 0.81 for sand, and 0.78 for silt.	LUCAS and KSSL	Limited to predicting soil texture properties (sand, silt, and clay)
[24]	Piccoli et al.	2023	Adaptive Deep Convolutional Neural Network. Neural Architecture Search (NAS).	Overall $R^2 \approx 0.75$.	LUCAS dataset. PRISMA simulated dataset.	relied mainly on hyperspectral information and did not include additional environmental or management factors.

[25]	Rahmani et al.	2022	Random Forest (RF), Cubist, Universal Kriging (UK)	SOM from 0.44 to 0.45, CEC from 0.39 to 0.44	(SSURGO)	limited to mapping two soil properties within a local study area and did not investigate multiple soil health indicators
[26]	Ramcharan et al.	2017	Random Forest	RMSPE was 0.13 gcm ⁻³ in average MPE was 0.001 gcm ⁻³ .	45,818 soils (NCSS)	focused on predicting soil bulk density only

References

- 1) F. A. Diaz-Gonzalez, J. Vuelvas, C. A. Correa, V. E. Vallejo, and D. Patino " Machine learning and remote sensing techniques applied to estimate soil indicators .. Review" 2021 Doi: 10.1016/j.ecolind.2021.108517
- 2) D.-C. Topa, S. Căpsună, A.-E. Calistru and C. Ailincăi" Sustainable Practices for Enhancing Soil Health and Crop Quality in Modern Agriculture _ A Review"2025 Doi:10.3390/agriculture15090998
- 3) J. B.-Valencia, V. G.-Pineda, E. R.-Vera, M. M.-Herrera, A. V.-Arias, R. H.-García, and J. Valencia" Machine Learning in Sustainable Agriculture _ Systematic Review and Research Perspectives"2025 Doi: 10.3390/agriculture15040377
- 4) S. Motia, and S. Reddy, “Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: a quantitative evaluation”2021 Doi: 10.1088/1742-6596/1950/1/012037
- 5) M. Awais, S. M. Z. A. Naqvi, H. Zhang, L. Li, W. Zhang, F. A. Awwad, E. A. A. Ismail, M. I. Khan, V. Raghavan and J. Hu*" AI and machine learning for soil analysis_an assessment of sustainable agricultural practices"2023 Doi: 10.1186/s40643-023-00710-y
- 6) T. B. Sheeba, L. D. V. Anand, G. Manohar, S. Selvan, B. Wilfred, K. Muthukumar, S. Padmavathy, P. R. Kumar, and B. T. Asfaw " Machine Learning Algorithm for Soil Analysis and Classification of Micronutrients in IoT-Enabled Automated Farms"2022 Doi:10.1155/2022/5343965
- 7) RV. R. Rao and U. S. Reddy “Agrimind intelligent deep learning framework for accurate soil fertility prediction” 2026 Doi: 10.1038/s41598-026-50366-9
- 8) H. Behera, R. K. Gouda,B. Nayak, N. Padhy, R. Panigrahi and P. K. Mahapatro “An In-Depth Comparative Analysis of Machine Learning Models for Soil Fertility Prediction” 2026 Doi: 10.3390/engproc2026124116
- 9) M. Venkatanaresh and I. Kullayamma “Deep learning based concurrent excited gated recurrent unit for crop recommendation based on soil and climatic conditions’ 2024 Doi: 10.1007/s11042- 023-18004-y

- 10) A. Badshah, F. Din, B. Y. Alkazemi, K. Z. Zamli, and M. Haris “Crop Classification and Yield Prediction Using Robust Machine Learning Models for Agricultural Sustainability” 2024 Doi: 10.1109/ACCESS.2024.3486653
- 11) A. Sarangi, S. K. Raula, S. Ghoshal, S. Kumar, S. Kumar, C. S. Kumar and N. Padhy “Enhancing Process Control in Agriculture: Leveraging Machine Learning for Soil Fertility Assessment” 2024 Doi: 10.3390/engproc2024067031
- 12) H. Chandra, P. M. Pawar, R. Elakkiya, P. S. Tamizharasan, R. Muthalagu, and A. Panthakkan” Explainable AI for Soil Fertility Prediction” 2023 Doi: 10.1109/ACCESS.2023.3311827
- 13) Z. Ghahremani, D. Huber, A. M. Rad, M. J. Johnson, and J. Pierce “A Machine Learning-Based High-Resolution Inventory of Soil Inorganic Carbon Across the Contiguous United States” 2026 Doi: 10.1029/2025GL120355
- 14) L. Sun, J. L. Safanelli, J. Sanderman, K. Georgiou, C. Brungard, K. Grover, B. G. Hopkins, Sh. Liu, and P.-T. Bremer “Self-supervised and multi-fidelity learning for extended predictive soil spectroscopy” 2026 Doi: 10.1016/j.geoderma.2026.117764
- 15) M. M. Khatun, Md. A. Hasan, Md S. A. Chyon, R. Anzoom, Md W. Ahmed, Md. S. Azam, S.M. H. S. Iqbal, Md. A. Rahim, Md. N. Imtiaz, and Md. T.Ahmed “Estimating soil organic carbon by integrating mid-infrared spectroscopy with explainable machine learning” 2026 Doi: 10.1016/j.rechem.2026.103386
- 16) Z. Aouabed, M. A. Bouaoune, V. Therrien, M. Bakhtyari, V. Makarenkov and M. Hijri “Soil microbiome prediction using traditional machine learning and deep learning models” 2026 Doi: 10.1038/s41598-026-39537-w
- 17) P. Kassai, G. Szatmári, M. Kocsis, B. Szabó, A. Makó, J. Mészáros, Z. Magyar, K. Takács, L. Pásztor, and A. Laborczi. “Large-scale mapping of soil particle size distribution using legacy data and machine learning-based pedotransfer functions” 2025 Doi: 10.1016/j.geoderma.2025.117178
- 18) L. Liu and X. Ma” Prediction of Soil Field Capacity and Permanent Wilting Point Using Accessible Parameters by Machine Learning” 2024 Doi: 10.3390/agriengineering6030151
- 19) T. W. Nauman, S. K.-Brown, S. M. Roecker, C. Brungard, D. White, J. Philippe, and J. A. Thompson “Soil landscapes of the United States (SOLUS): Developing predictive soil property maps of the conterminous United States using hybrid training sets” 2024 Doi: 10.1002/saj2.20769

- 20) Z. Zhai, H. Yu, F. Chen, J. Hu, H. Xu, and X. Zhou. "PS-MTL-LUCAS.. A partially shared multi-task learning model for simultaneously predicting multiple soil properties" 2024 Doi: 10.1016/j.econinf.2024.102784
- 21) M. M. Saberioon, A. Gholizadeh, A. Ghaznavi, S. Chabrilat, and V. Khosravi "Enhancing soil organic carbon prediction of LUCAS soil database using deep learning and deep feature selection" 2024 Doi: 10.1016/j.compag.2024.109494
- 22) B. Liu, B. Guo, R. Zhuo, F. Dai, and H. Chi "Prediction of the soil organic carbon in the LUCAS soil database based on spectral clustering" 2023 Doi: 10.17221/97/2022-SWR
- 23) O. P. Omondiagbe, Sh. A. Licorish, S. G. MacDonell, and L. Lilburne. "Soil texture prediction with automated deep convolutional neural networks and population-based learning" 2023 Doi: 10.1016/j.geoderma.2023.116521
- 24) F. Piccoli, R. Colombo, M. Rossini, P. Napoletano and R. Schettini "A deep scalable neural architecture for soil properties estimation from spectral information" 2023 Doi: 10.1016/j.cageo.2023.105433
- 25) Sh. R. Rahmani, J. P. Ackerson, D. Schulze, K. Adhikari and Z. Libohova "Digital Mapping of Soil Organic Matter and Cation Exchange Capacity in a Low Relief Landscape Using LiDAR Data" 2022 Doi: 10.3390/agronomy12061338
- 26) A. Ramcharan, T. Hengl, D. Beaudette, and S. Wills "A Soil Bulk Density Pedotransfer Function Based on Machine Learning: A Case Study with the NCSS Soil Characterization Database" 2017 Doi: 10.2136/sssaj2016.12.0421

