

الاء جاویش

thesisfinal.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:404412116

Submission Date

Aug 1, 2026, 2:25 PM GMT+3

Download Date

Aug 1, 2026, 7:54 PM GMT+3

File Name

thesisfinal.docx

File Size

4.5 MB

40 Pages

8,265 Words

50,222 Characters

45% Overall Similarity

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

Match Groups

-  **260 Not Cited or Quoted 35%**
Matches with neither in-text citation nor quotation marks
-  **21 Missing Quotations 3%**
Matches that are still very similar to source material
-  **31 Missing Citation 6%**
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

- 34%  Internet sources
- 38%  Publications
- 27%  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

- 260 Not Cited or Quoted 35%**
Matches with neither in-text citation nor quotation marks
- 21 Missing Quotations 3%**
Matches that are still very similar to source material
- 31 Missing Citation 6%**
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

- 34% Internet sources
- 38% Publications
- 27% 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	4%
2	Internet	doaj.org	3%
3	Internet	www.researchgate.net	2%
4	Internet	mdpi-res.com	2%
5	Publication	Most. Mira Khatun, Md. Abir Hasan, Md Sadiul Alam Chyon, Rashid Anzoom et al. ...	1%
6	Publication	Zahia Aouabed, Vincent Therrien, Mohamed Achraf Bouaoune, Mohammadreza B...	<1%
7	Publication	M. Venkatanaresh, I. Kullayamma. "Deep learning based concurrent excited gate...	<1%
8	Internet	ouci.dntb.gov.ua	<1%
9	Internet	assets-eu.researchsquare.com	<1%
10	Publication	RV. Raghavendra Rao, U. Srinivasulu Reddy. "Agrimind intelligent deep learning f...	<1%

11	Internet	experts.boisestate.edu	<1%
12	Internet	mila.quebec	<1%
13	Submitted works	Midocean University on 2025-09-05	<1%
14	Internet	www.kgeorgiou.com	<1%
15	Internet	journal.hep.com.cn	<1%
16	Internet	arxiv.org	<1%
17	Internet	www.jatit.org	<1%
18	Internet	acsess.onlinelibrary.wiley.com	<1%
19	Internet	scholar.archive.org	<1%
20	Publication	Luning Sun, José L. Safanelli, Jonathan Sanderman, Katerina Georgiou et al. "Self-...	<1%
21	Internet	www.ijraset.com	<1%
22	Internet	www.jetir.org	<1%
23	Publication	"Soils and Sustainable Agriculture", Springer Science and Business Media LLC, 2025	<1%
24	Internet	vectoral.org	<1%

25	Internet	www.thesesjournal.com	<1%
26	Internet	export.arxiv.org	<1%
27	Internet	www.gjet.edu	<1%
28	Publication	Zhenyuan Ji, Ao Yang, Zhiguo Liang, Haoxuan Yuan, Yun Zhang, Yong Du, Yaohua ...	<1%
29	Publication	S Krishnaveni, S Muthurajkumar. "Automated Machine Learning for Agricultural ...	<1%
30	Internet	masujournal.org	<1%
31	Internet	umpir.ump.edu.my	<1%
32	Submitted works	DIT University on 2026-06-26	<1%
33	Internet	dblp.org	<1%
34	Internet	ideas.repec.org	<1%
35	Publication	Jiaqi Shen, Peiwen Qin, Rui Zhong, Peiyao Han. "Optimization of Temporal Featur...	<1%
36	Internet	discovery.researcher.life	<1%
37	Publication	"Proceedings of Data Analytics and Management", Springer Science and Business...	<1%
38	Publication	Durgesh Kumar Mishra, Nilanjan Dey, Bharat Singh Deora, Amit Joshi. "ICT for Co...	<1%

39	Publication	Zhaoyu Zhai, Fuji Chen, Hongfeng Yu, Jun Hu, Xinfei Zhou, Huanliang Xu. "PS-MTL-...	<1%
40	Internet	mfuir.mfu.ac.th	<1%
41	Submitted works	Universiti Teknologi MARA on 2026-07-18	<1%
42	Submitted works	University of Arizona on 2024-06-05	<1%
43	Publication	Zahra Ghahremani, David Huber, Arash Modaresi Rad, Mike J. Johnson, Jennifer Pi...	<1%
44	Internet	dblp.dagstuhl.de	<1%
45	Internet	ijarcce.com	<1%
46	Publication	R.K. Gupta, R.K. Naresh, S.S. Walia, U.P. Singh, Dinesh Tomar, Ashok Kumar, Y.P. Si...	<1%
47	Publication	Preety Chaudhary, Preeti Gulia, Nasib Singh Gill. "An intelligent soil fertility evalu...	<1%
48	Submitted works	Vrije Universiteit Amsterdam on 2026-03-26	<1%
49	Internet	agrilandhub.com	<1%
50	Internet	www.clicours.com	<1%
51	Publication	Wenchao Qi, Le Yu, Weichao Sun, Tao Liu, Qiang Zhao, Yue Zhao, Hui Wu, Yibo Wa...	<1%
52	Internet	listens.online	<1%

53	Publication	Kanchanadevi P, Jayabharathy J. "Enhancing Soil Fertility Predictions with Sequen...	<1%
54	Submitted works	The University of Manchester on 2025-03-24	<1%
55	Internet	abusaxiy.uz	<1%
56	Internet	theses.lib.polyu.edu.hk	<1%
57	Internet	www.etd.ceu.edu	<1%
58	Internet	www.preprints.org	<1%
59	Internet	www.sciencepubco.com	<1%
60	Internet	harbinengineeringjournal.com	<1%
61	Internet	jisem-journal.com	<1%
62	Internet	www.ijert.org	<1%
63	Publication	Jabulani Jolly Matsimbe, Megersa Olumana Dinka, David Olatunde Olukanni. "Sus...	<1%
64	Publication	Muresan, Anca Olivia. "Graph-Based and Uncertainty-Aware Machine Learning fo...	<1%
65	Submitted works	Nottingham Trent University on 2007-03-16	<1%
66	Submitted works	University of East London on 2026-05-07	<1%

67	Internet	mSPACE.lib.umanitoba.ca	<1%
68	Internet	swr.agriculturejournals.cz	<1%
69	Internet	hal.science	<1%
70	Publication	"Intelligent Vision and Computing", Springer Science and Business Media LLC, 2026	<1%
71	Submitted works	Agriculture & Forestry University on 2026-07-22	<1%
72	Submitted works	Bahrain Polytechnic on 2023-05-30	<1%
73	Submitted works	Graphic Era University on 2025-07-18	<1%
74	Publication	Shankar Babu, Mahesh Babu Kota. "Synergies in Smart and Virtual Systems using..."	<1%
75	Internet	dokumen.pub	<1%
76	Internet	gfzpublic.gfz.de	<1%
77	Internet	journal.idnns.org	<1%
78	Internet	neptjournal.com	<1%
79	Submitted works	Galgotias University, Greater Noida on 2026-06-14	<1%
80	Internet	boa.unimib.it	<1%

81	Internet	www.bartleby.com	<1%
82	Internet	www.cedengineering.com	<1%
83	Internet	www.iis.fraunhofer.de	<1%
84	Submitted works	College of Public Health on 2026-01-29	<1%
85	Submitted works	KJSCE on 2026-07-08	<1%
86	Publication	Liu Hai, Song Yong, Du Qingfu. "Power Forecasting of Combined Heating and Coo...	<1%
87	Submitted works	Unity College on 2026-07-27	<1%
88	Submitted works	University of Glasgow on 2026-04-25	<1%
89	Submitted works	University of Stirling on 2010-04-19	<1%
90	Internet	archive.org	<1%
91	Internet	www.hindawi.com	<1%
92	Publication	Braimah, Mohammed. "Spectral Decomposition and Model Estimation of Soil Car...	<1%
93	Publication	Debasis Chaudhuri, Jan Harm C Pretorius, Debashis Das, Sauvik Bal. "Internationa...	<1%
94	Submitted works	ICTS on 2026-07-07	<1%

95	Submitted works	Midocean University on 2026-05-02	<1%
96	Internet	federicagazzelloni.com	<1%
97	Internet	jeas.springeropen.com	<1%
98	Internet	kilimo.go.ke	<1%
99	Internet	wefnexus.tamu.edu	<1%
100	Internet	www.bestpfe.com	<1%
101	Submitted works	Associatie K.U.Leuven on 2026-05-09	<1%
102	Publication	Dothang Truong. "Data Science and Machine Learning for Non-Programmers - Us...	<1%
103	Publication	Pritha Kundu, Saikat Biswas. "Securing soil health as the foundation for strength...	<1%
104	Publication	Rajesh Bhanudas Shinde, N. T. Markad, Mangal V. Patil. "Chapter 36 Soil Property ...	<1%
105	Publication	Reza Yousefzadeh, Alireza Kazemi, Rashid S. Al-Maamari. "Application of power-la...	<1%
106	Submitted works	The Independent Institute of Education (IIE) on 2025-11-14	<1%
107	Submitted works	University of Derby on 2026-03-31	<1%
108	Submitted works	University of Sheffield on 2024-09-10	<1%

109	Submitted works	University of Stellenbosch, South Africa on 2024-11-01	<1%
110	Internet	etda.libraries.psu.edu	<1%
111	Internet	society.org	<1%
112	Internet	www.alkej.uobaghdad.edu.iq	<1%
113	Internet	www.old-aj.cz	<1%
114	Internet	www.researchsquare.com	<1%
115	Submitted works	Bournemouth University on 2023-08-27	<1%
116	Publication	Giannis Gallios, Nikolaos Tsakiridis, Nikolaos Tziolas. "Federated learning applicat...	<1%
117	Submitted works	Islington College,Nepal on 2026-07-30	<1%
118	Submitted works	Kingston University on 2025-08-25	<1%
119	Publication	Saman Rabiei, Ebrahim Babaeian, Sabine Grunwald. "Deep learning-based downs...	<1%
120	Submitted works	University of Edinburgh on 2011-08-18	<1%
121	Submitted works	University of Sunderland on 2026-07-02	<1%
122	Internet	cgspace.cgiar.org	<1%

123	Internet	ddescholar.acemap.info	<1%
124	Internet	frontierus.org	<1%
125	Internet	geneticsmr.com	<1%
126	Internet	inria.hal.science	<1%
127	Internet	pub.epsilon.slu.se	<1%
128	Internet	www.numberanalytics.com	<1%
129	Internet	5dok.net	<1%
130	Publication	Akhtar Hameed, Muhammad Zahid. "Soil Nanotechnology", CRC Press, 2026	<1%
131	Submitted works	Anna University on 2024-08-09	<1%
132	Submitted works	Anna University on 2026-03-31	<1%
133	Publication	Aryal, Rakshya. "Climate Smart Agricultural Practices and Their Effects on Soil Art..."	<1%
134	Submitted works	British University in Egypt on 2025-06-13	<1%
135	Submitted works	Coventry University on 2024-09-11	<1%
136	Submitted works	Cranfield University on 2011-04-29	<1%

137	Publication	Dinesh Goyal, Bhanu Pratap, Sandeep Gupta, Saurabh Raj, Rekha Rani Agrawal, I...	<1%
138	Submitted works	Higher Education Commission Pakistan on 2010-02-20	<1%
139	Publication	Poonam Nandal, Tapas Kumar, Meeta Singh, Mamta Dahiya. "Progressive Compu...	<1%
140	Publication	Sujata Dash, Subhendu Kumar Pani, Joel J. P. C. Rodrigues, Babita Majhi. "Deep Le...	<1%
141	Submitted works	Universitas Esa Unggul on 2026-07-20	<1%
142	Submitted works	University of Sheffield on 2021-03-29	<1%
143	Publication	Wenqi Guo, Shichen Gao, Yaohui Ding, Daming Dong. "Simultaneous prediction o...	<1%
144	Internet	csitic.com	<1%
145	Internet	fastercapital.com	<1%
146	Internet	hseb-conferences.kz	<1%
147	Submitted works	i-CATS University College on 2026-06-08	<1%
148	Internet	ijirt.org	<1%
149	Internet	joiv.org	<1%
150	Internet	mzuir.inflibnet.ac.in	<1%

151	Internet	onclick360.com	<1%
152	Internet	pdfs.semanticscholar.org	<1%
153	Submitted works	universiteteknologimara on 2025-07-09	<1%
154	Internet	www.iaeng.org	<1%
155	Internet	www.landcan.org	<1%
156	Internet	www.rbcjournal.org	<1%
157	Internet	www.s-tronomic.in	<1%
158	Publication	"AI and Data Analytics in Precision Agriculture for Sustainable Development", Spr...	<1%
159	Publication	"Computational Intelligence in Internet of Agricultural Things", Springer Science ...	<1%
160	Publication	Chengcheng Xu, Jingyi Huang, Alfred E. Hartemink, Nathaniel W. Chaney. "Prune...	<1%
161	Publication	Chunming Chi, Lili Yang, Long Ma, Cuili Zhang. "Utilizing Hygroscopic Water Cont...	<1%
162	Publication	David Olatunde Olukanni, Oluseyi Olanrewaju Ajayi, Joshua Olusegun Okeniyi, Sh...	<1%
163	Publication	Florin Nenciu, Mihai Gabriel Matache, Iuliana Gageanu, Ioan Catalin Persu, Florin ...	<1%
164	Publication	Gusmawartati, Via Permata Sari, M Joehari Jamili. " Effect of Green Manure from ...	<1%

165	Publication	Harvinder Singh, Priyanka Kaushal, Sarabjeet Kaur. "Advanced Computing and AI...	<1%
166	Submitted works	Mepco Schlenk Engineering college on 2024-04-21	<1%
167	Publication	Nokwazi Zanele Ngubo, Zama Eric Mashimbye, Kyle Loggenberg. "The impact of s...	<1%
168	Publication	Pattwat Maneewattanapinyo, Chaowalit Monton, Jirapornchai Suksaeree. "Advan...	<1%
169	Submitted works	Scotland's Rural College on 2023-03-07	<1%
170	Submitted works	University of Arid Agriculture, Murree Road on 2026-07-16	<1%
171	Publication	Amanda Ramcharan, Tomislav Hengl, Dylan Beaudette, Skye Wills. "A Soil Bulk De...	<1%
172	Submitted works	Asian Institute of Technology on 2026-04-20	<1%
173	Publication	Baoyang Liu, Baofeng Guo, Renxiong Zhuo, Fan Dai, Haoyu Chi. "Prediction ofthe ...	<1%
174	Publication	C. Channarayappa, D.P. Biradar. " Soil Basics, Management, and Rhizospher...	<1%
175	Publication	Dothang Truong. "Demystifying AI - Data Science and Machine Learning Using IB...	<1%
176	Publication	Guangyao Sun, Wencai Zhang, Yuanhao Lou, Tonglin Liao et al. "Soil knowledge-g...	<1%
177	Publication	Halbast Rashid Ismail, Gheyath Mustafa Zebari, Ibrahim Mahmood Ibrahim, Adel ...	<1%
178	Publication	Harmesh Behera, Bibhukalyan Nayak, Ritesh Kumar Gouda, Neelamadhab Padhy,...	<1%

179	Publication	Harshiv Chandra, Pranav M. Pawar, R. Elakkiya, P S Tamizharasan, Raja Muthalag...	<1%
180	Publication	Juan Botero-Valencia, Vanessa García-Pineda, Alejandro Valencia-Arias, Jackeline ...	<1%
181	Publication	Osayande Pascal Omondiagbe, Linda Lilburne, Sherlock A. Licorish, Stephen G. M...	<1%
182	Publication	Subhabrata Panda. "Agroforestry", Springer Science and Business Media LLC, 2025	<1%
183	Submitted works	Teaching and Learning with Technology on 2024-07-27	<1%
184	Submitted works	Vrije Universiteit Amsterdam on 2026-03-26	<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 of human population, in correlation the need for food increases making new issues emerge due to climate change and 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 plant growth, nutrient availability, water retention, and ecosystem stability. Traditional techniques for evaluating soil health are frequently expensive, time-consuming, 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 with machine learning techniques 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.

Contents

Acknowledgements		1
Abstract		2
Chapter 1: Introduction		4
1.1 Overview		4
1.2 Problem Definition		5
1.3 Thesis Objectives		6
1.4 Thesis Contribution		7
1.5 Thesis Organization		7
Chapter 2: Background and Related Work		8
2.1 Introduction		8
2.2 Background		9
2.3 Related work		22
References		35

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 one of the most crucial aspects to ensure the productivity of agriculture, as it affects various processes such as crop growth, nutrient availability, water retention and ecosystem stability. Hence soil health is a key ingredient of sustainable agriculture over the years. 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 is the capacity of soil to sustain 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].

The advent of Artificial Intelligence (AI) and Machine Learning (ML) has opened up new possibilities for soil analysis and agricultural management [3]. Machine learning algorithms can be used to analyze extensive agricultural data, uncover intricate connections between soil markers, and make precise soil condition predictions [1]. Models like Random Forest, Support Vector Machines, and

Neural Networks have demonstrated potential success in soil data analysis and in decision-making processes related to agriculture [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 various physical, chemical and biotic properties of soil [2].

This research aims to examine the relationships between the various soil health indicators, determine the most significant indicators and facilitate sustainable agricultural management practices by applying machine learning analysis to soil health indicators.

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 various physical, chemical and biological measurements. While important, these methods also are time-consuming, expensive, and hard to apply on a continuing basis over large agricultural areas [1].

Soil health is also impacted by several indicator factors working together with soil moisture, pH, organic matter, nutrient availability and microbial activity [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].

Although several studies have been performed on the applications of machine learning in agriculture, most of them have been developed to address a single soil property and not for the simultaneous evaluation of several soil health indicators. 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.

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. In this chapter we will show some of the most important concepts of soil health, soil health indicators, and agriculture sustainability. In addition to the machine learning methods that have been used for soil analysis and agriculture decision-making.

Soil fertility management is considered one of the most important factors for maintaining sustainable agricultural production. It helps the farmers to maintain the ability of the soil to provide the appropriate amount of nutrients to the plants when it is needed. Machine learning techniques have been applied for soil analysis recently. 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, this chapter discusses the background concepts of soil health, agricultural sustainability, and machine learning techniques that are used in agriculture. 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 plays a significant role in supporting plants growth, maintains environmental 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 helping mitigate climate change.

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

Soil health is influenced by many physical, chemical, and biological factors. Physical properties include soil structure, soil texture, and water holding capacity. The chemical properties include soil pH, potential pollutants, organic carbon, nutrient availability, and electrical conductivity. Biological properties also play an important role 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 combination of different indicators to provide a more comprehensive assessment [2].

Soil health is evaluated using a set of indicators including the physical, chemical, and biological condition of the soil, which are used to provide information about the soil ability to support plants growth and maintain ecosystem services. Different indicators are combined to provide accurate assessment of soil condition [2].

Sustainable agriculture aims to meet the growing demand for food while protecting natural resources. It focuses on improving agricultural productivity

through the efficient use of soil and other resources without compromising the needs of future generations. Sustainable agriculture now has progressively adopted advanced technologies. These technologies include precision agriculture, remote sensing, and the internet of things (IOT), to improve farm management and support informed decision making [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 sensors, drones, satellites, and other smart devices. Traditional methods are no longer sufficient for analyzing these data. Thus, machine learning has emerged as an effective approach for processing agriculture data, identifying hidden patterns, and supporting accurate and timely decision making in sustainable agriculture [3].

Several machine learning algorithms have been applied in agricultural research to analyze complex datasets and improve prediction accuracy. These algorithms examples are Decision Tree, Random Forest, Support Vector Machine, multilayer perceptron (MLP) neural networks, Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN). These algorithms have been used in different agricultural applications, such as soil analysis, crop yield prediction, disease detection, and precision farming. These algorithms have the ability to analyze large datasets and identify complex relationships, making them valuable tools for supporting sustainable agriculture management [3]. In addition, the combination of ML with the IoT is revolutionizing smart agriculture. In recent years, researchers have investigated the use of these technologies to automate agricultural operations, such as monitoring soil conditions, weather, irrigation management, and pest control [3].

2.2.1 Soil Health Indicators

soil health indicators are measurable soil properties The indicators used to evaluate the condition of soil and help evaluate its ability to support plant growth, maintain environmental quality. These indicators provide valuable information about the physical, chemical, and biological characteristics of soil. To assess soil condition, a combination of different indicators is used. [2]

2.2.2 Physical Indicators

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

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

2- Soil Texture. Soil texture describes the organization of soil particles and the spaces that exist between them. [1], [2]

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

4- Soil Structure. Soil structure refers to the arrangement of soil particles into aggregates that improves water infiltration, aeration, root penetration, 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 such as soil pH, soil organic carbon, nitrogen, phosphorus, potassium, and electrical conductivity (EC) influence plant growth, and microbial activity, influence plant growth, and microbial activity.[1][2]

1- Soil pH. Soil pH measures the acidity or alkalinity of soil. Extreme pH values can reduce nutrient uptake and negatively influence crop productivity. [2]

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

3- Nitrogen (N). Nitrogen is an essential nutrient it plays a key role in protein synthesis and photosynthesis. 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]

164
49
5- **Potassium (K).** Potassium is an essential nutrient that regulates several processes in plants, including water balance, enzyme activation, and resistance to environmental stress.[2]

90
174
6- **Electrical Conductivity (EC).** Electrical conductivity (EC) measures the concentration of soluble salts in the soil and is used to assess soil salinity which can affect the volume of water uptake by plants.[2]

2.2.4 Biological Indicators

170
75
Biological indicators evaluate the living components and provide information about microbial activity. Common biological indicators include soil organic matter, microbial activity, and soil respiration. [1], [2].

1- **Soil Organic Matter (SOM).** Soil organic matter consists of decomposed plant and animal residues present in the soil and it improves soil structure.

169
2- **Microbial Activity.** Microbial activity refers to the activity of microorganisms such as bacteria and fungi in the soil high microbial activity mean healthy soil.[2]

72
3- **Soil Respiration.** Soil respiration measures the amount of carbon dioxide released by soil microorganisms during the decomposition of organic matter.

2.2.5 Machine Learning Algorithms

58
4
The management of agricultural systems has been changed because of the rapid development of digital technologies. Modern farming generates large amounts of data from sensors, drones, satellites, and other smart devices. Traditional methods are no longer sufficient for analyzing these data. Thus, machine learning has emerged as an effective approach for processing agriculture data, identifying hidden patterns, and supporting accurate and timely decision making in sustainable agriculture [3].

107
25
Several machine learning (ML) algorithms have been applied in agricultural research to analyze complex datasets and improve prediction accuracy. These algorithms examples are Decision Tree, Random Forest, Support Vector Machine, multilayer perceptron (MLP) neural networks, Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN).

41
These algorithms have been used in different agricultural applications, such as soil analysis, crop yield prediction, disease detection, and precision farming.

These algorithms have the ability to analyze large datasets and identify complex relationships, making them valuable tools for supporting sustainable agriculture management [3]. In addition, the combination of ML with the IoT is revolutionizing smart agriculture. In recent years, researchers have investigated the use of these technologies to automate agricultural operations, such as monitoring soil conditions, weather, irrigation management, and pest control [3].

1. Random Forest (RF)

is an ensemble algorithm that combines the results of several decision trees in order to improve the accuracy of data classification and predictions the algorithm is particularly reliable since it allows to work with noisy data. prevents overfitting and is effective for both large amounts of data and a high number of input variables.

Regression results from using an ensemble algorithm can be calculated using the following formula:

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

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

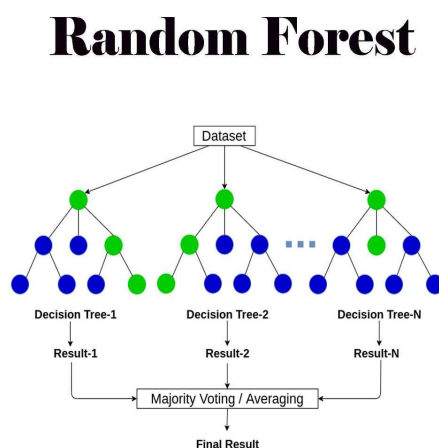


figure 2.1

2. Support Vector Machine (SVM)

is a supervised machine learning algorithm that is used for classification and regression analysis. It is based on finding the optimal separating hyperplane that divides the space of classes. SVM uses kernel functions for the best approximation of the surface between the categories. In the case of a linear model,

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.

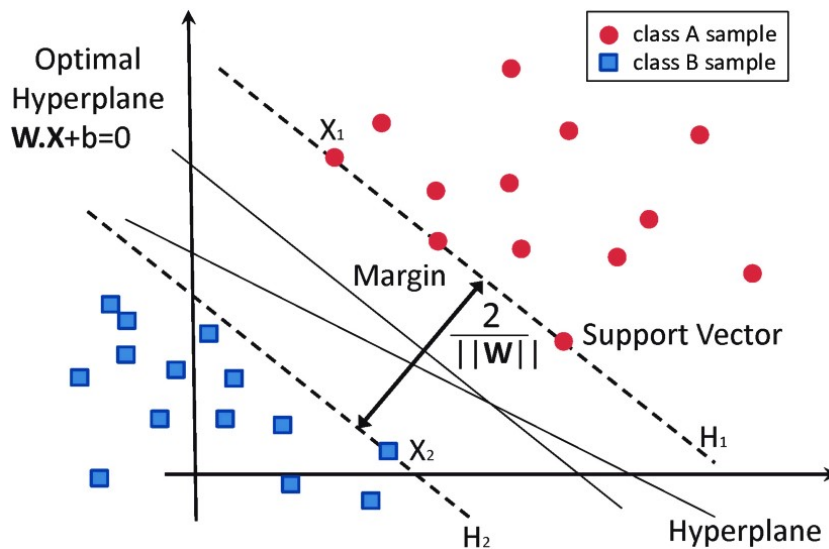


figure 2.2

3. Decision Tree (DT).

is a supervised learning algorithm 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 c is the total number of classes and p_i is the probability of an item belonging to class i .

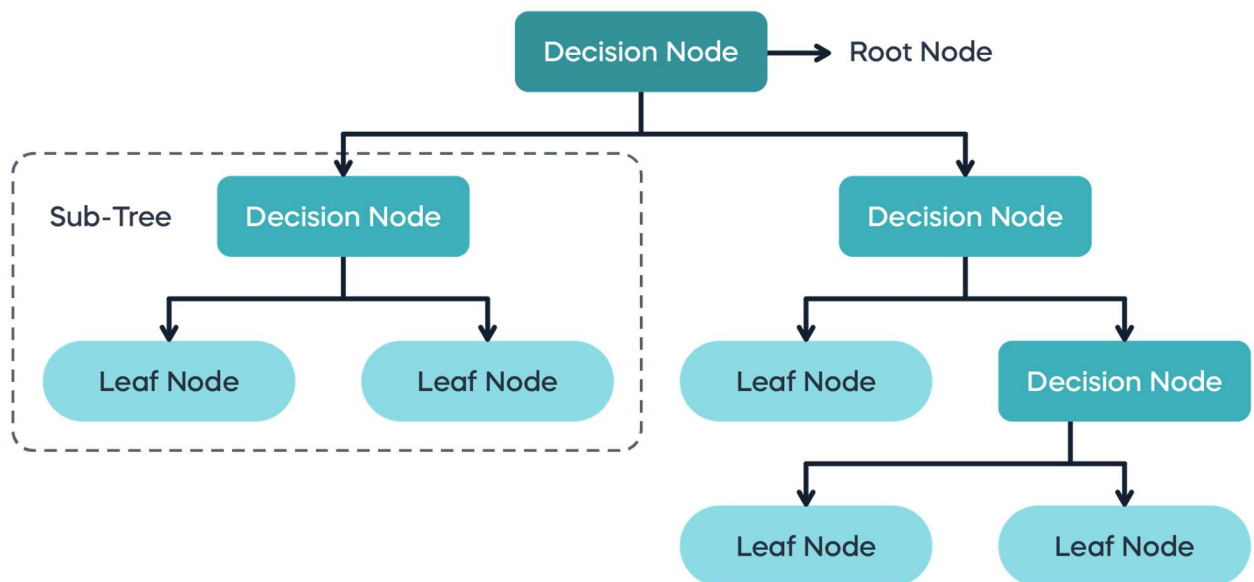


Figure 2.3

4. Artificial Neural Network (ANN).

is a computing system which is designed after the human brain. It is an artificial neural network made up of input, hidden, and output layers. the artificial neural network ANN is very effective at capturing the complex nonlinear relationships of different physical and chemical properties of soil health indicators. The output of an individual neuron is given by the following equation:

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

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

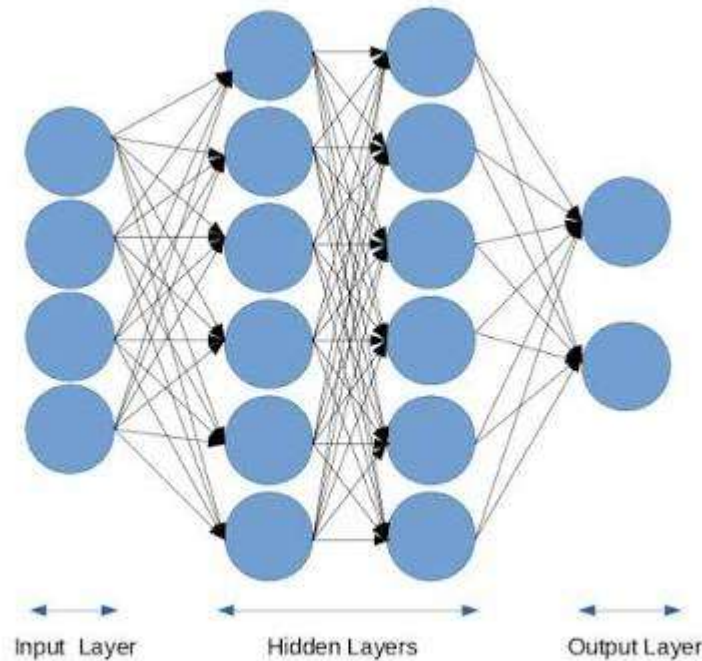


Figure 2.4

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.

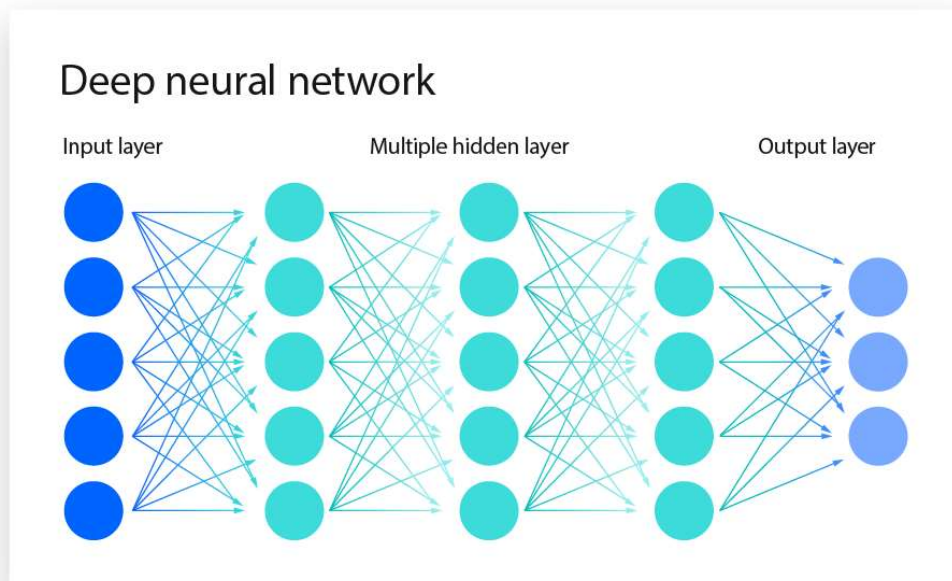


Figure 2.5

6. Extreme Learning Machine (ELM).

Extreme Learning Machine is a feedforward neural network where weights of the hidden layer are randomly assigned and fixed and only the weights of the output layer are calculated 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:

$$\beta = H^+ Y \quad (6)$$

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

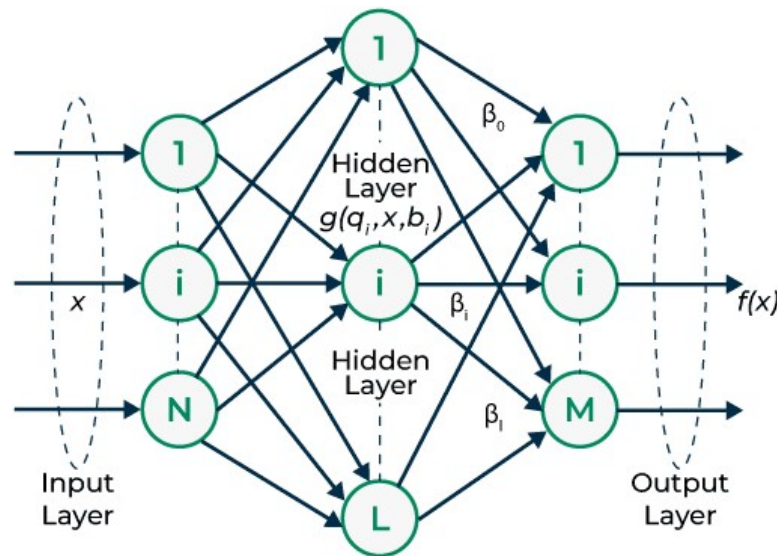


Figure 2.6

7. Gradient Boosting Machine (GBM).

Gradient Boosting Machine builds trees sequentially where each subsequent tree is trained to correct the errors of the previous one. In contrast to the random forest algorithm, gradient boosting 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 minimizing the loss function:

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

(7)

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

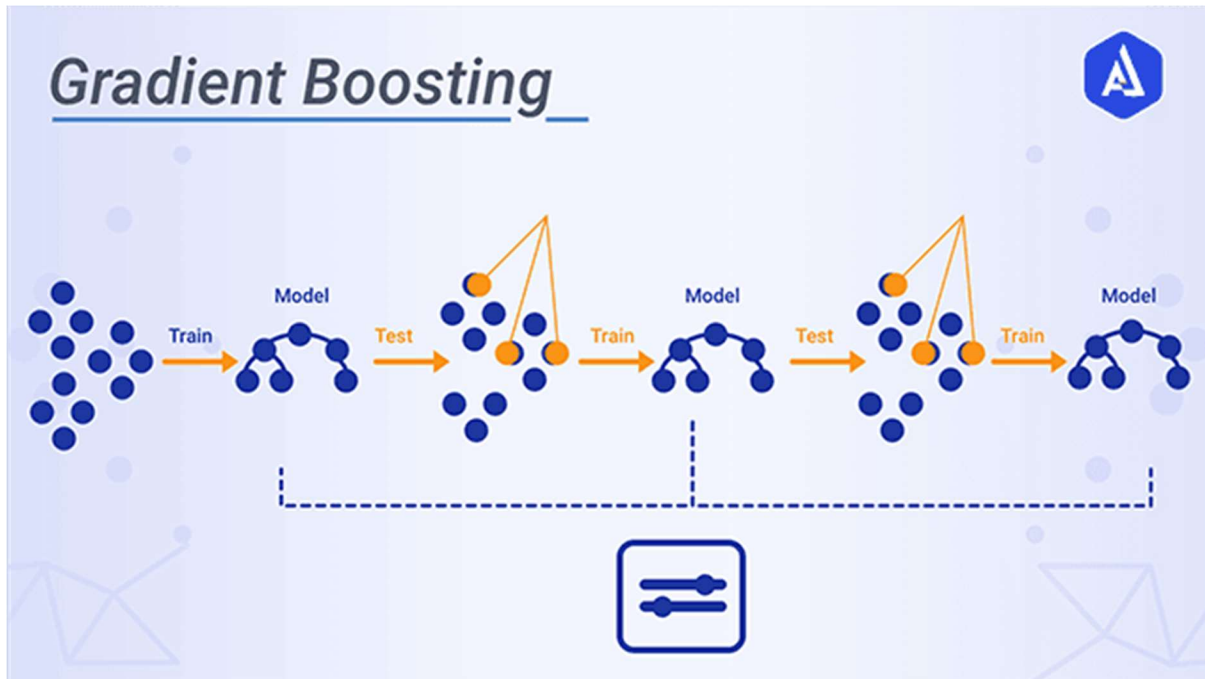


Figure 2.7

8. XGBoost.

XGBoost is an optimized regularized implementation 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 is the loss function f_t is the new tree and $\Omega(f_t)$ penalizes tree complexity.

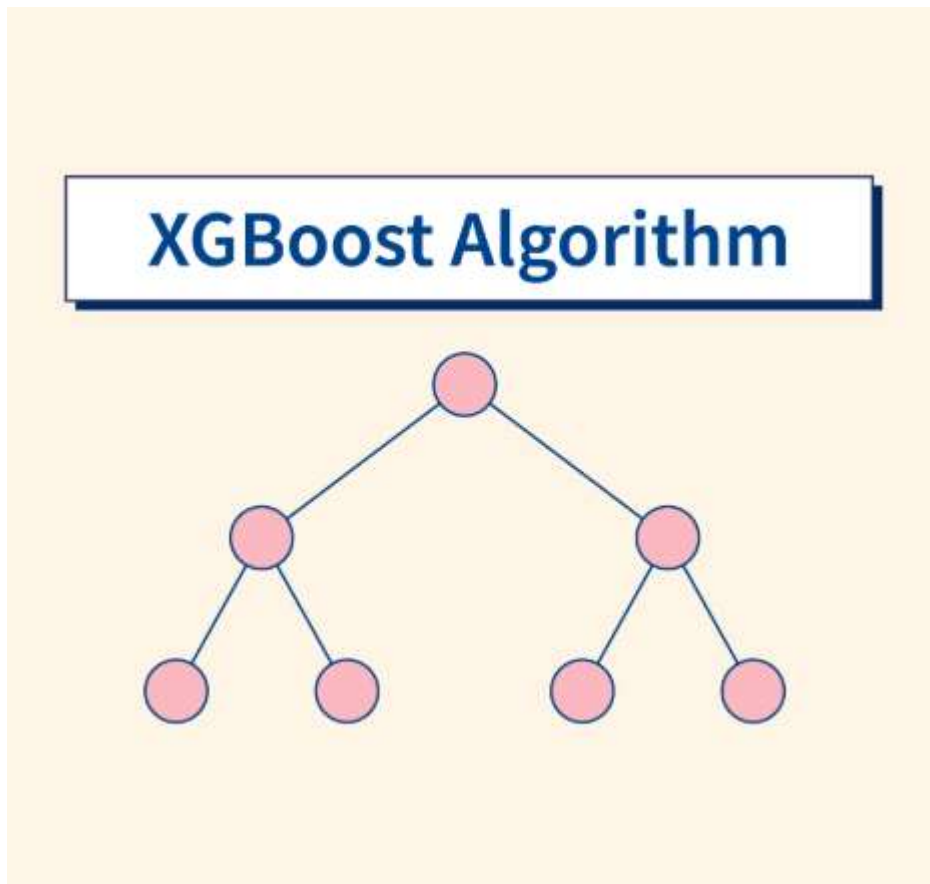


Figure 2.8

9. Convolutional Neural Network (CNN)

Convolutional Neural Network is a deep learning model used to process data structured 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:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n) \quad (9)$$

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

2.2.6 Evaluation Metrics

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

- **Coefficient of Determination (R^2):** Evaluates how well 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): Measures the square root of the average of the squared differences between the predicted and actual values. It is highly sensitive to large errors.

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

Mean Absolute Error (MAE): Measures the average of the absolute differences between the predicted values and the actual values, providing a linear assessment 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:

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

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:

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

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.

Rao and Reddy (2026) proposed the AgriMind Intelligent Fertility Predictor (AMIFP) for soil fertility classification using deep learning techniques. The study used a publicly available Kaggle dataset containing 880 soil samples. The proposed framework integrated a novel preprocessing method (SIC), an attention-based deep learning model (ALADN), and a genetic hyperparameter optimization algorithm (GPSHO). The dataset was partitioned into 70% for training, 10% for validation, and 20% for held-out testing, ensuring an unbiased evaluation on unseen data. The model achieved 98.65% accuracy, 98.73% precision, 98.65% recall, and 98.70% F1-score, demonstrating high performance in soil fertility classification. Although the proposed framework achieved excellent classification performance, its effectiveness was validated using only one publicly available dataset.[7]

Behera et al. (2026) conducted a comparative study to evaluate the performance of multiple machine learning classifiers for soil fertility prediction. The study used a publicly available Kaggle dataset containing 880 soil samples described by 12 physicochemical soil properties. This study used 16 machine learning classifiers for soil fertility prediction, such as learner-based, ensemble-based,

instance-based, and probabilistic-based models. The model's performance is assessed using accuracy, precision, recall, and F1-score. The results showed that the RF model achieves an accuracy of 90.91% compared to the other models. However, the model showed relatively low performance for one of the minority classes, indicating challenges associated with class imbalance and overlapping feature distributions. [8]

Venkatanaresh and Kullayamma (2024) proposed a novel deep learning (DL) technique, Concurrent Excited Gated Recurrent Unit (CEGRU), for automatic crop recommendation based on soil, fertilizer, and climatic conditions. The study used publicly available Kaggle datasets and real-time collected Maharashtra datasets. To improve the model performance, the Hunter-Prey Optimization (HPO) algorithm was employed for parameter tuning. The proposed method was evaluated using several performance metrics, including accuracy, kappa, positive predictive value (PPV), recall, sensitivity, RMSE, and time complexity. The result showed that the overall accuracy of 99.2% and 99.3%, PPV of 97.5% and 98.3%, kappa of 95.7% and 96.6%, RMSE of 1.05 and 1.50, the time complexity of 105 s is obtained for both Maharashtra and Kaggle datasets respectively. However, the study focused on recommending suitable crops based on soil, fertilizer, and climatic conditions, while the proposed research focuses on analyzing multiple soil health indicators.[9]

Badshah et al. (2024) proposed two robust machine learning models for classification and regression based on two different datasets. The first was a crop recommendation dataset obtained from Kaggle, which included various input attributes such as soil pH, temperature, humidity, and nutrient levels. Several machine learning classification algorithms were evaluated, including Extra Tree Classifier (ETC), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM). The results showed that the Random Forest (RF) model achieved the best classification performance with an accuracy of 99.7% with precision, recall, F1 score, and confusion matrix by using K-fold cross-validation, Explainable AI (XAI) and feature engineering. The second was obtained from World Bank and Food and Agriculture Organization (FAO) dataset, used for wheat production and forecast the yield using machine learning regression

models. The result showed that hyper parameter tuning with K-fold cross-validation, Support Vector Regressor (SVR) achieved best performance with an accuracy of 99.9% with R2 Score. In addition, Explainable AI techniques, including Feature Importance and LIME, were applied to interpret model predictions. 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]

Sarang et al. (2024) developed a machine learning model for soil fertility assessment. The model was based on soil properties like N, P, K, Ph, nutrient level, moisture levels, temp, rainfall, and topography. The study used a publicly available dataset from Kaggle. The objective of the study was to classify soil into fertile or non-fertile. Four machine learning classifiers are used to determine the best classifier based on the performance metrics including Logistic Regression (LR), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Decision Tree (DT). The results showed 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 the implementation of an explainable AI (XAI) model for soil fertility prediction by using Random Forest classification. In this study the LUCAS 2018 Topsoil dataset was used. It used several soil parameters such as pH, organic carbon, nitrogen, electrical conductivity, phosphorus, and potassium. The study combined k- Means clustering, Random Forest, and SHAP to explain the result. The result showed an accuracy of 97.02%, which shows the effectiveness of the explainable machine learning for soil fertility assessment. However, this study focused on soil fertility prediction only, and did not analyze multiple soil health indicators [12].

Ghahremani et al. (2026) introduced a high-resolution (30 m) CONUS SIC map using machine learning (ML) models trained on the ISRIC World Soil Information Service (WoSIS) database together with climate and land surface variables. Two machine learning models, Random Forest Regression (RFR) and Random Forest Classification (RFC), were employed to predict Soil Inorganic Carbon (SIC) distribution. The results showed that the Random Forest Regression model achieved the best performance with an RMSE of 15 kg/m², while the Random Forest Classification model achieved a classification accuracy of 0.56. The study also identifies soil pH as the most influential predictor in the formation of these carbon-rich soils. However, the study focused on mapping soil inorganic carbon while the proposed research aims to analyze multiple soil health indicators to better understand their contribution to sustainable agriculture.[13]

Sun et al. (2026) proposed a self-supervised multi-fidelity learning framework for predicting multiple soil properties from spectral information. The study used the data that sourced from the USDA Natural Resources Conservation Service National Soil Survey Center Kellogg Soil Survey Laboratory (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 Partial Least Squares Regression (PLSR) and conventional MLP models. The results showed that the MIR-SSL-MLP model achieved the best performance for most soil properties, obtaining R² values 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) explored the use of spectroscopy integrated with machine learning as a non-destructive and efficient alternative to predict soil organic carbon (SOC) based on mid-infrared spectroscopy (MIR) data within the spectral range of 600–4000 cm⁻¹. The study utilized the publicly available Open Soil Spectral Library (OSSL), which integrates soil spectral data from several international databases, including KSSL, LUCAS, RaCA, AfSIS, and ICRAF–ISRIC. Partial Least Squares Regression (PLSR) and Random Forest (RF) models

were developed to predict SOC from MIR spectral data, while SHAP (Shapley Additive Explanations) was employed to interpret the contribution of individual spectral wavelengths. The RF model was optimized through hyperparameter tuning using the Grid-Search method. The RF model achieved the best performance with a coefficient of determination (R²) value of 0.98 and a root mean square error (RMSEP) of 1.45%. 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) focused on estimating the composition of bacterial and fungal communities at the functional and taxonomic scales by analyzing the data from two publicly available soil microbiome datasets: (1) Data collected by Averill et al., and (2) Data extracted from the NEON database using traditional machine learning and deep learning models. Several machine learning models, including Random Forest (RF), Gradient Boosting (GB), k-Nearest Neighbors (k-NN), Support Vector Regression (SVR), Ridge Regression, Lasso Regression, and Multilayer Perceptron (MLP), were evaluated and compared. The experimental results showed that traditional machine learning models generally outperformed the deep learning model across most prediction tasks, particularly when training data were limited. The best overall results obtained include the value of the coefficient of determination $R^2=0.57$ at the phylum taxonomic level, provided by the Gradient boosting model for the bacterial dataset, while value of $R^2=0.45$ at the functional group level, provided by both the Random forest and Gradient boosting models for the fungal dataset. 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) developed pedotransfer functions (PTFs) to compute the particle size distribution from the soil properties. The study utilized the publicly available NATASA database, which contains 1,372 soil profiles, while the Hungarian Detailed Soil Hydrophysical Database, comprising 3,970 soil profiles, was used to train and evaluate the prediction models by using the random forest

method. The Random Forest Kriging (RFK) was used to generate high-resolution digital soil maps for sand, silt, and clay fractions. The results showed that the R^2 values of the PTFs are 0.69 for clay, 0.58 for silt, and 0.74 for sand content. In addition, the RFK models achieved R^2 values ranging from 0.19 to 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 the WoSIS database. The study used artificial neural network (ANN) and gene-expression programming (GEP) algorithms to predict the soil hydraulic properties. The results showed that the ANN model achieved the best performance for PWP prediction with a normalized root mean square error (NRMSE) of 19.9%, while the GEP model performed best for FC prediction with an NRMSE of 29.9%. In addition, the mean absolute error (MAE) was reduced by 51.54% for the FC and 56.38% for the PWP by the ANN model, compared with the conventional linear models. 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 Cooperative Soil Survey (NCSS) Soil Characterization Database, the Soil Survey Geographic Database (SSURGO), and the National Soil Information System (NASIS), to generate 100 m resolution maps for 20 soil properties, including various textural fractions, physical parameters, chemical parameters, carbon, and depth to restrictions. A hybrid training strategy helped increase training data by roughly 10-fold over previous similar studies. 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). The study employed a multi-task deep learning architecture combined with SHapley Additive exPlanations (SHAP) to improve prediction performance and interpret important spectral features. LUCAS Topsoil Dataset (2009) was used to predict seven soil properties, including pH, nitrogen (N), phosphorus (P), potassium (K), cation exchange capacity (CEC), organic carbon (OC), and calcium carbonate (CaCO_3). The results showed high prediction performance, achieving R^2 values of 0.945, 0.936, 0.413, 0.624, 0.837, 0.952, and 0.956 for pH, N, P, K, CEC, OC, and CaCO_3 , respectively. However, the model was evaluated 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) focused on Enhancing soil organic carbon prediction by using deep learning techniques. The study used 11441 samples collected from LUCAS Soil Database. The study compared several models, including One-Dimensional Convolutional Neural Network (1DCNN), Fully Connected Neural Network (FCNN), Stacked AutoEncoder (SAE), and Random Forest (RF) model was implemented for comparison. The results showed that the SAE-1DCNN model achieved the best prediction performance with an R^2 value of 0.784 and an RMSE of 3.94. However, the study focused only on predicting Soil Organic Carbon (SOC) and did not analyze multiple soil health indicators affecting agricultural sustainability.[21]

Liu et al. (2023) investigated the prediction of Soil Organic Carbon (SOC) by using the Visible and Near Infrared Diffuse Reflectance Spectroscopy (Vis-NIR DRS) tool. The study used the LUCAS EU-wide Topsoil Database for comparing several prediction approaches, including Global Partial Least Squares Regression (PLSR), Local PLSR, k-means clustering with PLSR, and hierarchical clustering with PLSR. The results showed that the best results were obtained for the local PLSR approach with $R^2 = 0.75$, root mean squared error of prediction (RMSEP) = 13.38 g/kg and ratio of performance to interquartile range (RPIQ) = 2.846, but the algorithm running time was 30.05 s. Similar results were obtained for the k-means clustering method with $R^2 = 0.75$, RMSEP = 14.61 g/kg and RPIQ =

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) proposed an automated deep learning model for soil texture prediction. The study used Deep Convolutional Neural Networks (CNNs) with Adapted Population-Based Training (Adapted-PBT) and Bayesian Optimization to improve model performance. LUCAS Soil Library and the Kellogg Soil Survey Laboratory (KSSL) Used as a dataset. The result showed that the PBT model achieved the best prediction performance, with model efficiency coefficient (MEC) values of 0.90 for clay, 0.81 for sand, and 0.78 for silt. However, the study focused 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 relies on a flexible neural architecture able to automatically adapt to the spectral library during analysis and allows for backtracking the spectral bands that most contribute to the prediction of a given variable. The results showed that the model achieves an overall R^2 of ~ 0.75 . In detail, 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. All models performed similarly for both SOM and CEC. With R^2 values ranging 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 method to develop a soil bulk density pedotransfer function (PTF) using the Random Forest. The study used 45,818 soil horizons extracted from the publicly available National Cooperative Soil Survey (NCSS) Soil Characterization Database in the United States. The data included surficial materials and hierarchical ecosystem land classifications. The results of a five-fold cross-validation showed that the model root mean squared prediction error (RMSPE) was 0.13 g cm⁻³ in average, while the mean prediction error (MPE) was -0.001 g cm⁻³. In addition, the PTF were used with soil organic carbon (SOC) stocks to fill in data gaps for volumetric assessments. However, The study focused on predicting soil bulk density only and did not analyze multiple soil health indicators or evaluate their relationships. [26]

Ref	Author	Year	Technique Used	Accuracy	Dataset	Comparison with This Research
[7]	Rao and Reddy	2026	SIC ALADN GPSHO	98.65% accuracy, 98.73% precision, 98.65% recall, and 98.70% F1- score	Kaggle dataset containing 880 soil samples	using only one publicly available dataset
[8]	Behera et al.	2026	16 machine learning classifiers as learner-based, ensemble-based, instance-based, and probabilistic- based models	RF model achieves an accuracy of 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 + Maharasht ra	focused only on recommending suitable crops based on soil, fertilizer
[10]	Badshah et al	2024	Extra Tree Classifier (ETC), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K- Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM).	best performance by Random Forest with an accuracy of 99.7% using K-fold cross- validation.	crop recommen dation dataset obtained from Kaggle, FAO	study focused mainly on crop recommendation and yield prediction
[11]	Sarangi et al.	2024	Logistic Regression (LR), K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (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 Informatio n 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 soil organic carbon from spectral information
[16]	Aouabed et al.	2026	Random Forest (RF) Gradient Boosting (GB) k-Nearest Neighbors (k-NN) Support Vector Regression (SVR) Ridge Regression Lasso Regression Multilayer Perceptron (MLP) Regressor	The best overall R ² =0.57 at the phylum taxonomic level by (GB)	Averill et al. Dataset NEON Dataset	study only 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	Artificial Neural Network (ANN) Gene Expression Programming (GEP) Linear Regression for comparison	ANN model achieved the best for PWP with NRMSE of 19.9% while GEP with an NRMSE of 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, 0.936, 0.413, 0.624, 0.837, 0.952, and 0.956 for pH, N, P, K, CEC, OC, and CaCO ₃ , respectively.	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 model achieved the best performance with an R ² of 0.784 and an RMSE of 3.94.	11441 samples collected from LUCAS Soil Database	Limited to predicting Soil Organic Carbon (SOC)
[22]	Liu et al.	2023	Global PLSR Local PLSR k-means clustering PLSR Hierarchical clustering PLSR	best results for local PLSR with R ² = 0.75, (RMSEP) = 13.38 g/kg RPIQ = 2.846	LUCAS EU-wide Topsoil Database	focused only on Soil Organic Carbon prediction
[23]	Omondiagbe et al.	2023	(Adapted-PBT) Bayesian Optimization (CNN)	PBT best prediction with (MEC) f 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 g cm^{-3} in average MPE was -0.001 g cm^{-3} .	45,818 soil (NCSS)	focused on predicting soil bulk density only

References

- 1) Freddy A. Diaz-Gonzalez, Jose Vuelvas, Carlos A. Correa b, Victoria E. Vallejo c, D. Patino " Machine learning and remote sensing techniques applied to estimate soil indicators – Review" 2021 <https://doi.org/10.1016/j.ecolind.2021.108517>
- 2) Denis-Constantin Topa, Sorin Căpsună, Anca-Elena Calistru and Costică Ailincăi "Sustainable Practices for Enhancing Soil Health and Crop Quality in Modern Agriculture: A Review" 2025 <https://doi.org/10.3390/agriculture15090998>
- 3) Juan Botero-Valencia, Vanessa García-Pineda, Erick Reyes-Vera, Mateo Mejia-Herrera, Alejandro Valencia-Arias, Ruber Hernández-García, and Jackeline Valencia " Machine Learning in Sustainable Agriculture: Systematic Review and Research Perspectives" 2025 <https://doi.org/10.3390/agriculture15040377>
- 4) Sanjay Motia, and SRN Reddy, "Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: a quantitative evaluation" 2021 <https://doi.org/10.1088/1742-6596/1950/1/012037>
- 5) Muhammad Awais, Syed Muhammad Zaigham Abbas Naqvi, Hao Zhang, Linze Li, Wei Zhang, Fuad A. Awwad, Emad A. A. Ismail, M. Ijaz Khan, Vijaya Raghavan and Jiandong Hu* " AI and machine learning for soil analysis: an assessment of sustainable agricultural practices" 2023 <https://doi.org/10.1186/s40643-023-00710-y>
- 6) T. Blesslin Sheeba, L. D. Vijay Anand, Gunaselvi Manohar, Saravana Selvan, Bazil Wilfred, K. Muthukumar, S. Padmavathy, P. Ramesh Kumar, and Belete Tessema Asfaw " Machine Learning Algorithm for Soil Analysis and Classification of Micronutrients in IoT-Enabled Automated Farms" 2022 <https://doi.org/10.1155/2022/5343965>
- 7) RV. Raghavendra Rao & U. Srinivasulu Reddy "Agrimind intelligent deep learning framework for accurate soil fertility prediction" 2026 <https://doi.org/10.1038/s41598-026-50366-9>
- 8) Harmesh Behera, Bibhukalyan Nayak, Ritesh Kumar Gouda, Neelamadhab Padhy, Rasmita Panigrahi and Pradeep Kumar Mahapatro "An In-Depth Comparative

Analysis of Machine Learning Models for Soil Fertility Prediction” 2026
<https://sciforum.net/event/ASEC2025>

- 9) M. Venkatanaresh & I. Kullayamma “Deep learning based concurrent excited gated recurrent unit for crop recommendation based on soil and climatic conditions’ 2024
<https://doi.org/10.1007/s11042-023-18004-y>

- 10) Abid Badshah; Basem Yousef Alkazemi; Fakhrud Din; Kamal Z. Zamli; Muhammad Haris “Crop Classification and Yield Prediction Using Robust Machine Learning Models for Agricultural Sustainability” 2024
<https://doi.org/10.1109/ACCESS.2024.3486653>

- 11) Ashutosh Sarangi, Sailesh Kumar Raula, Sohamdev Ghoshal, Swadhin Kumar, Swadhin Kumar, Chinta Sai Kumar and Neelamadhab Padhy “Enhancing Process Control in Agriculture: Leveraging Machine Learning for Soil Fertility Assessment” 2024 <https://doi.org/10.3390/engproc2024067031>

- 12) HARSHIV CHANCHANDRA, NAV M. PAWAR ELAKKIYA R., TAMIZHARASAN P S, RAJA MUTHALAGU AND ALAVIKUNHU PANTHAKKAN” Explainable AI for Soil Fertility Prediction” 2023<https://doi.org/10.1109/ACCESS.2023.3311827>

- 13) Zahra Ghahremani, David Huber, Arash Modaresi Rad, Mike J. Johnson, Jennifer Pierce “A Machine Learning-Based High-Resolution Inventory of Soil Inorganic Carbon Across the Contiguous United States” 2026
<https://doi.org/10.1029/2025GL120355>

- 14) Luning Sun, José L. Safanelli, Jonathan Sanderman, Katerina Georgiou, Colby Brungard, Kanchan Grover, Bryan G. Hopkins, Shusen Liu, Peer-Timo Bremer “Self-supervised and multi-fidelity learning for extended predictive soil spectroscopy” 2026 <https://doi.org/10.1016/j.geoderma.2026.117764>

- 15) Most. Mira Khatun, Md. Abir Hasan, Md Sadiul Alam Chyon, Rashid Anzoom, Md Wadud Ahmed, Md. Shafiul Azam, S.M. Hasan Sazzad Iqbal, Md. Abdur Rahim, Md. Niaz Imtiaz, Md. Toukir Ahmed “Estimating soil organic carbon by integrating mid-infrared spectroscopy with explainable machine learning” 2026
<https://doi.org/10.1016/j.rechem.2026.103386>

- 16) Zahia Aouabed, Vincent Therrien, Mohamed Achraf Bouaoune, Mohammadreza Bakhtyari, Mohamed Hijri and Vladimir Makarenkov “Soil microbiome prediction using

traditional machine learning and deep learning models” 2026

<https://doi.org/10.1038/s41598-026-39537-w>

- 17) Piroska Kassai, Mihály Kocsis, Gábor Szatmári, András Makó, János Mészáros, Annamária Laborczi, Zoltán Magyar, Katalin Takács, László Pásztor, Brigitta Szabó “Large-scale mapping of soil particle size distribution using legacy data and machine learning-based pedotransfer functions” 2025
<https://doi.org/10.1016/j.geoderma.2025.117178>

- 18) Liwei Liu and Xingmao Ma”Prediction of Soil Field Capacity and Permanent Wilting Point Using Accessible Parameters by Machine Learning” 2024
<https://doi.org/10.3390/agriengineering6030151>

- 19) Travis W. Nauman, Suzann Kienast-Brown, Stephen M. Roecker, Colby Brungard, David White, Jessica Philippe, James A. Thompson “Soil landscapes of the United States (SOLUS): Developing predictive soil property maps of the conterminous United States using hybrid training sets” 2024 <https://doi.org/10.1002/saj2.20769>

- 20) Zhaoyu Zhai, Fuji Chen, Hongfeng Yu, Jun Hu, Xinfei Zhou, Huanliang Xu. ”PS-MTL-LUCAS: A partially shared multi-task learning model for simultaneously predicting multiple soil properties” 2024 <https://doi.org/10.1016/j.ecoinf.2024.102784>

- 21) Mohammad Mehdi Saberioon, Asa Gholizadeh, Ali Ghaznavi, Sabine Chabrilat, Vahid Khosravi “Enhancing soil organic carbon prediction of LUCAS soil database using deep learning and deep feature selection”2024
<https://doi.org/10.1016/j.compag.2024.109494>

- 22) Baoyang Liu, Baofeng Guo, Renxiong Zhuo, Fan Dai, Haoyu Chi “Prediction of the soil organic carbon in the LUCAS soil database based on spectral clustering” 2023
<https://doi.org/10.17221/97/2022-SWR>

- 23) Osayande Pascal Omondiagbe, Linda Lilburne, Sherlock A. Licorish, Stephen G. MacDonell.” Soil texture prediction with automated deep convolutional neural networks and population-based learning” 2023 <https://doi.org/10.1016/j.geoderma.2023.116521>

- 24) Flavio Piccoli, Micol Rossini, Roberto Colombo, Raimondo Schettini, Paolo Napoletano “A deep scalable neural architecture for soil properties estimation from spectral information” 2023 <https://doi.org/10.1016/j.cageo.2023.105433>

- 25) Shams R. Rahmani, Jason P. Ackerson, Darrell Schulze, Kabindra Adhikari and Zamir Libohova “Digital Mapping of Soil Organic Matter and Cation Exchange Capacity in a Low Relief Landscape Using LiDAR Data” 2022
<https://doi.org/10.3390/agronomy12061338>

 18

 18

- 26) Amanda Ramcharan, Tomislav Hengl, Dylan Beaudette, Skye Wills "A Soil Bulk Density Pedotransfer Function Based on Machine Learning: A Case Study with the NCSS Soil Characterization Database" 2017
<https://doi.org/10.2136/sssai2016.12.0421>

