SPECTRAL-BASED SOIL NUTRIENT PREDICTION USING MACHINE LEARNING TECHNIQUES

A Thesis Submitted
in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

Submitted by

SHAGUN JAIN

Roll No. 2K22/PHDCS/02

Under the supervision of
Dr. Divyashikha Sethia (Supervisor)
Associate Professor
Department of Software Engineering
and
Dr. (Col) Kailash Chandra Tiwari (Co-Supervisor)
Professor
Department of Geospatial Science and Technology



To the

Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India.

October, 2026

Acknowledgement

I want to express my sincere gratitude to my supervisor, Dr. Divyashikha Sethia, Associate Professor, Department of Software Engineering, and co-supervisor, Dr. (Col) Kailash Chandra Tiwari, Professor, Department of Geospatial Science and Technology, for their motivation, counselling, constant support, and encouragement, which have helped me in formulating an urge for this research thesis. Their exemplary hard work and insightful guidance have always helped me in overcoming difficulties.

My sincere regards to Prof. Prateek Sharma, Vice-Chancellor, Delhi Technological University, for providing a platform to pursue my PhD work. I express my thankfulness to Prof. Ruchika Malhotra, Head, Department of Software Engineering, for her kind support and for providing the necessary facilities to undertake this research.

I sincerely thank Dr. Shalini Gakhar, Data Scientist at IRRI, and Dr. Rabi N. Sahoo, Principal Scientist at ICAR, Delhi, for allowing me to work at ICAR and gain valuable experience in data collection and remote sensing methodologies relevant to my research.

I am deeply grateful to my husband, Adv. Vaibhav Jain, for his constant encouragement and support. I sincerely thank my parents, Mr. Sanjay Jain and Mrs. Bindu Jain, my sister, CA Sakshi Jain, my brother, Akshit Jain, all my brothers-in-law, and my sisters-in-law, for their love and support. I also extend my heartfelt gratitude to my mother-in-law, Judge (Retd.) Mrs. Sharda Jain, for her blessings and encouragement.

Also, a huge thanks to all my friends who have continuously supported and helped me with their valuable suggestions as and when required. Finally, I would like to acknowledge everyone who, in one way or another, contributed to the successful completion of this work. Your encouragement and support, whether big or small, have helped me reach this significant milestone in my life. Thank you all.

Shagun Jain

2K22/PHDCS/02



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I **Shagun Jain** hereby certify that the work which is being presented in the thesis entitled **"SPECTRAL-BASED SOIL NUTRIENT PREDICTION USING MACHINE LEARNING TECHNIQUES"** in partial fulfillment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the Department of **Software Engineering**, Delhi Technological University is an authentic record of my own work carried out under the supervision of **Dr. Divyashikha Sethia**, Associate Professor, Department of Software Engineering, DTU, and **Dr. (Col) Kailash Chandra Tiwari**, Professor, Department of Geospatial Science and Technology, DTU.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Shagun Jain
2K22/PHDCS/02

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s) **Signature of External Examiner**



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

This to certify that **Shagun Jain** (2K22/PHDCS/02) has carried out her research work presented in this thesis entitled **"SPECTRAL-BASED SOIL NUTRIENT PREDICTION USING MACHINE LEARNING TECHNIQUES"** for the award of **Doctor of Philosophy** from Department of Software Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Supervisor	**Co-Supervisor**
Dr. Divyashikha Sethia	Dr. (Col) Kailash Chandra Tiwari
Associate Professor	Professor
Department of Software Engineering	Department of Geospatial Science and Technology
Delhi Technological University	Delhi Technological University

Date:

Indian Examiner

ABSTRACT

Achieving the United Nations' Sustainable Development Goal (SDG) of Zero Hunger requires enhancing agricultural productivity and sustainable soil management practices. Soil nutrients are essential for crop growth, but conventional laboratory-based chemical analyses, though accurate, are slow, costly, labor-intensive, and unsuitable for large-scale or frequent monitoring. These constraints limit the timely assessment of soil health, especially in remote or resource-limited regions. Advances in Artificial Intelligence (AI) have enabled the use of low-cost RGB (Red, Green, and Blue) imaging for soil analysis. However, its three broad bands capture limited spectral information, making it suitable for general soil classification rather than precise nutrient prediction. To overcome these limitations, remote sensing—particularly multispectral and hyperspectral imaging—has emerged as a powerful alternative. These sensors provide rich spectral signatures across multiple wavelengths, enabling non-invasive, rapid, and scalable prediction of soil nutrients. When combined with Machine Learning (ML) and Deep Learning (DL) models that can capture complex soil–spectral relationships, spectral imaging offers a promising path toward efficient, accurate, and large-scale soil nutrient monitoring for modern agriculture.

This thesis aims to improve the automation and accuracy of soil nutrient prediction by leveraging machine learning, deep learning, and spectral index engineering. Unlike existing approaches that often overlook optimal feature selection or index design, the proposed methodology focuses on extracting meaningful spectral features and developing tailored predictive models suited for complex spectral patterns. This thesis addresses key research gaps, including: (1) inadequate research on selecting the most informative spectral band subsets for nutrient prediction; (2) limited application of DL models optimized for complex spectral patterns; (3) underexplored development

of novel spectral indices for nutrient-specific prediction; and (4) the lack of studies applying existing spectral indices to hyperspectral datasets for soil nutrient analysis. Addressing these challenges helps in the development of scalable, automated, and accurate solutions for soil nutrient monitoring.

Firstly, this thesis presents a critical systematic review of AI-based soil nutrient prediction using spectral data, offering statistical insights into model performance, spectral preprocessing, and dataset usage. It highlights ML/DL applications, feature selection, and spectral indices while identifying open challenges, research gaps, and publicly available datasets, and providing recommendations and future research opportunities.

Secondly, this thesis conducts a comparative evaluation of various feature selection techniques—including all features, Principal Component Analysis (PCA), Least Absolute Shrinkage and Selection Operator (LASSO), Successive Projections Algorithm (SPA), and Recursive Feature Elimination (RFE) —to identify the most effective method for modeling different soil nutrients. It helps enhance model performance by selecting the most relevant spectral bands for each nutrient.

Thirdly, this thesis explores the comparative efficacy of various ML and DL models. It proposes a hybrid ML model (HyPLS-GPR) utilizing multispectral data, as well as two attention-based frameworks employing hyperspectral data for soil nutrient prediction, namely CNViS-Net and CAT-MTLNet. A Partial Least Squares Regression (PLSR) – Gaussian Process Regression (GPR) hybrid model (HyPLSR-GPR) using multispectral Landsat-8 data effectively predicts multiple nutrients across Belgium and Luxembourg, demonstrating acceptable performance with an RMSE range of 0.92 to 0.98 for various soil nutrients. Meanwhile, CNViS-Net, a self-attention framework that combines a Convolutional Neural Network (CNN) and a Vision-Transformer (ViT) with Support Vector Regression (SVR), demonstrates state-of-the-art accuracy on the HYPERVIEW dataset, achieving remarkably low RMSE in the range of 0.23 to 21.91 for phosphorus, potassium, magnesium, and pH prediction. This thesis also introduces CAT-MTLNet, a novel deep learning framework designed for multi-target soil nutrient prediction using hyperspectral data (LUCAS-2009 and AfSIS). The CATMTL-Net framework combines the strengths of two powerful components: a Concrete Autoencoder (CAE) for data-driven spectral feature selection and a Transformer-MLP-based network for

learning complex, long-range dependencies within the spectral signatures. The CAE automatically identifies the most informative wavelengths by enforcing feature sparsity through concrete gating, resulting in a compact and interpretable subset of spectral bands. These selected features are then processed by the Transformer-MLP, whose self-attention mechanism effectively models global relationships across wavelengths and captures subtle variations linked to multiple soil nutrients. By jointly learning and predicting several soil nutrients within a multi-target learning framework, CAT-MTLNet enhances both accuracy (RMSE in the range of 0.289 to 39.006) and generalization. It enables an efficient, robust, and scalable approach to hyperspectral soil analysis as compared to traditional single-target and manually engineered feature selection methods.

Lastly, the work addresses the lack of nutrient-specific spectral indices by proposing two novel spectral indices, the Soil Organic Carbon Index (SOCI) and the Soil pH Index (SPI), for an improved prediction of soil pH and organic carbon using LUCAS-2009 and AfSIS hyperspectral datasets. These indices are developed using Principal Component Analysis (PCA) and an Artificial Neural Network (ANN). PCA selects the most informative spectral wavelengths, which form the basis of the index structure. The ANN then uses these PCA-selected wavelengths as inputs to formulate the final index equations for predicting SOC and pH, ensuring effective feature selection. Experimental results, with RMSE values ranging from 0.28 to 19.62, demonstrate that these indices consistently outperform traditional band-based inputs. Additionally, a comparative analysis of existing indices, including vegetation, soil, water, temperature, and thermal indices, reveals their relative effectiveness in predicting soil pH and organic carbon levels.

Collectively, this thesis bridges key gaps in AI-enabled soil nutrient prediction by integrating systematic evaluation, algorithmic innovation, and domain-specific spectral understanding. It further paves the way for precision agriculture solutions that are robust, interpretable, and scalable.

List of Publications

Papers Published in International Journals

Journal 1: S. Jain, D. Sethia and K.C. Tiwari. "A Critical Systematic Review on Spectral-Based Soil Nutrient Prediction Using Machine Learning." **Environmental Monitoring and Assessment** 196, 699 (July 2024).(**SCIE, Impact Factor: 3.6, Publisher: Springer**).

Journal 2: S. Jain, D. Sethia and K.C. Tiwari. "Developing Novel Spectral Indices for Precise Estimation of Soil pH and Organic Carbon With Hyperspectral Data and Machine Learning." **Environmental Monitoring and Assessment**: 196, 1255 (November 2024). (**SCIE, Impact factor-3.6, Publisher: Springer**).

Journal 3: S. Jain, D. Sethia and K.C. Tiwari. "CAT-MTLNet: A Novel Deep Multi-Target Framework for Hyperspectral Soil Properties Estimation Using Concrete Autoencoder and Transformer" **IEEE Geoscience and Remote Sensing Letters** (November 2025). **(SCIE, Impact factor-4.8, Publisher: IEEE**).

Journal 4: S. Jain, D. Sethia and K.C. Tiwari. "TLM-Stack: A Deep Learning-based Novel Framework for Soil Nutrients Estimation Using Hyperspectral Data" **Computers and Electronics in Agriculture** (March 2026). **(SCIE, Impact factor-10.3, Publisher: Elsevier)**.

Papers Published in International Conferences

Conference 1: S. Jain, and D. Sethia. "A Review on Applications of Artificial Intelligence for Identifying Soil Nutrients." in Springer *Proc.* First International Conference on Agriculture-Centric Computation, May 2023.

Conference 2: **S. Jain**, D. Sethia and K.C. Tiwari. "A Hybrid Approach for Soil Nutrient Estimation using Multispectral Data in Belgium and Luxembourg." in IEEE *Proc.* First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI), October 2023.

Conference 3: **S. Jain**, D. Sethia and K.C. Tiwari. "Comparative Analysis of Machine Learning Based Soil pH Prediction Using Spectral Bands and Indices." in Springer *Proc.* Second International Conference on Agriculture-Centric Computation, (pp. 43-55), February 2025.

Conference 4: **S. Jain**, D. Sethia and K.C. Tiwari. "CNViS-Net: A Self-Attention Framework for Predicting Soil Nutrients from Hyperspectral Data." in IEEE *Proc.* International Conference on Frontier Technologies and Solutions, March 2025.

Conference 5: A. Shah, S. Sah, S. Singhal, **S. Jain** and D. Sethia "DSeP-xNet: A Feature Optimized Ensemble Framework for Hyperspectral-Based Soil Organic Carbon Prediction." in IEEE *Proc.* International Conference on Augmented Reality, Intelligent System, and Industrial Automation, December 2024.

Conference 6: A. Sangwan, D. Sethia and **S. Jain** "Multi-Strategy Fused Transformer-CNN with PSO-Driven Optimization for Soil Properties Estimation from Hyperspectral Data" in IEEE *Proc.* International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), March 2026.

Contents

ABSTRACT i

List of Publications iv

List of Tables xii

List of Figures xv

List of Abbreviations xviii

Chapter 1: Introduction 1

1.1 Methods to Predict Soil Nutrients 3

1.1.1 Laboratory-based Chemical Analysis: 3

1.1.2 Imaging-based soil nutrient prediction: 5

1.1.2.1 RGB Imaging: 5

1.1.2.2 Spectral Imaging: 7

1.2 Motivation and Background 9

1.2.1 Why Remote Sensing for Soil Nutrient Prediction? 9

1.2.2 Electromagnetic Spectrum 9

1.2.2.1 Electromagnetic Spectrum Components and Applications 10

1.2.2.2 Key Concepts in Remote Sensing for Soil Nutrient Prediction 12

1.2.3 Spectral Data and Spectral Imaging 13

1.2.4 Types of Spectral Data: Multispectral and Hyperspectral . . . 13

1.3 Automated Spectral-based Machine Learning for Soil Nutrient Prediction 14

1.3.1 Research Gaps 15

1.4 Problem Statement 17
1.5 Research Objectives and Contribution of the Thesis 17
1.6 Significance of Thesis 19
1.7 Thesis Overview 20

Chapter 2: Technical Background of Remote Sensing for Soil Nutrient Prediction 23

2.1 Remote Sensing 23
2.1.1 Types of Remote Sensing 24
2.1.2 Applications of Remote Sensing 25
2.1.3 Electromagnetic Spectrum 27
2.1.3.1 Relevant Spectral Regions in Remote Sensing: . . . 28
2.1.3.2 What are Spectral Bands? 28
2.1.3.3 What is Spectral Reflectance? 29
2.1.3.4 What are the Spectral Signatures of Earth Surface Materials? 30
2.1.4 What is Spectral Data? 30
2.1.4.1 Significance of Spectral Reflectance in Remote Sensing 32
2.1.5 What is Spectral Imaging? 33
2.1.6 Acquisition of Spectral Data and Images: Sensors and Platforms 34
2.1.6.1 Calibration and Correction 39
2.2 Types of Spectral Data 40
2.2.1 Multispectral Data 41
2.2.1.1 Key Characteristics 41
2.2.2 Hyperspectral Data 42
2.2.2.1 Key Characteristics 43
2.2.3 Difference between Hyperspectral and Multispectral Data . . . 43
2.3 Advantages of Spectral data 46
2.4 Challenges in Spectral Data Analysis 47
2.5 Automated Spectral Data Preprocessing Pipeline for Machine Learning Based Prediction 48

2.5.1 Conversion of Raw Spectral Images into Spectral Reflectance Values . . . 49
2.5.2 Noise Removal . . . 51
2.5.2.1 Review and Discussion on Noise Removal Techniques 54
2.5.3 Spectral Normalization . . . 55
2.5.4 Feature Selection . . . 55
2.6 Summary . . . 56

Chapter 3: Literature Review **58**
3.1 Open Spectral Soil Datasets . . . 58
3.1.1 A Database of Land Use and Coverage Area frame Survey (LUCAS-2009) [1] . . . 60
3.1.2 LUCAS-2015 Topsoil Dataset with Landsat-8 Imagery [2] . . 62
3.1.3 A Database of Africa Soil Information Service (AfSIS) [3] . . 65
3.1.4 HYPERVIEW Hyperspectral Soil Dataset [4] . . . 66
3.1.5 Rupnagar Soil Dataset with Landsat-8 Imagery [5] . . . 69
3.1.6 Univalle Hyperspectral Soil Phosphorus Dataset [6] . . . 71
3.2 Performance Metrics . . . 73
3.3 Feature Selection Techniques for Soil Nutrient Prediction using Spectral Data . . . 74
3.4 Classification of ML and DL Models . . . 79
3.4.1 List of ML Models . . . 80
3.4.2 List of DL Models . . . 86
3.4.3 Contribution of ML Techniques in Soil Nutrient Prediction . . 90
3.4.4 Contribution of DL Techniques in Soil Nutrient Prediction . . 91
3.5 Spectral Indices for Soil Nutrient Prediction . . . 92
3.6 Limitations of Existing Machine Learning in Soil Nutrient Prediction . 97
3.7 Summary . . . 99

Chapter 4: HyPLS-GPR: A Hybrid Machine Learning Model for Soil Nutrient Prediction Using Multispectral Data **101**
4.1 Motivation . . . 101
4.2 Related Work . . . 103

4.3 Experimental Methodology 104
4.3.1 Dataset Description 105
4.3.2 Data Preprocessing 107
4.3.2.1 Feature Engineering 109
4.3.3 Proposed HyPLS-GPR model 110
4.3.3.1 Parameter Tuning and Regression Algorithms 110
4.3.4 Performace Metrics 112
4.4 Results and Discussion 113
4.4.1 Individual Model Performance on LUCAS-2015 [2] 114
4.4.2 Rationale of the Proposed HyPLS-GPR Model 115
4.4.3 Validation of HyPLS-GPR on Rupnagar Dataset [5] 117
4.4.4 Discussion 118
4.5 Summary 118

Chapter 5: CNViS-Net: A Self-Attention Framework for Soil Nutrient Prediction Using Hyperspectral Data **120**

5.1 Motivation 120
5.2 Related Work 122
5.3 Experimental Methodology 123
5.3.1 Dataset Description 123
5.3.2 Feature Extraction 124
5.3.3 Architecture of Proposed CNViS-Net Framework 124
5.3.4 Parameter Tuning and Regression Algorithms 126
5.3.5 Performance Metrics 127
5.4 Results and Discussion 127
5.4.1 Individual Model Performance on HYPERVIEW [4] 128
5.4.2 Rationale of the Proposed CNViS-Net Framework 132
5.4.3 Validation of CNViS-Net on Univalle Dataset [6] 133
5.4.4 Discussion 133
5.5 Summary 134

Chapter 6: CAT-MTLNet: A Novel Deep Multi-Target Framework for Hyperspectral Soil Nutrient Prediction Using Concrete Autoencoder and Transformer **136**

6.1 Motivation . . . 137
6.2 Related Work . . . 138
6.3 Experimental Methodology . . . 139
6.3.1 Dataset Description . . . 139
6.3.2 Data Preprocessing . . . 140
6.3.3 Model Architecture . . . 142
6.3.3.1 Transformer Encoder Block . . . 142
6.3.3.2 MLP Regression Head . . . 142
6.3.4 Evaluation of Alternative Feature Selection Techniques . . . 143
6.3.5 Evaluation Metrics . . . 143
6.4 Results and Discussion . . . 144
6.4.1 Experimental Results and Analysis . . . 144
6.4.1.1 Evaluation of proposed CAT-MTLNet Performance . 144
6.4.1.2 Comparison of Feature Selection Techniques . . . 145
6.4.1.3 Analysis of Selected Spectral Bands . . . 146
6.4.1.4 Experimental Comparison with Existing Models . . 147
6.4.2 Discussion . . . 148
6.5 Summary . . . 150

Chapter 7: Development of Novel Spectral Indices for Soil pH and Organic Carbon Prediction **151**

7.1 Motivation . . . 151
7.2 Related Work . . . 153
7.3 Experiment Methodology . . . 155
7.3.1 Dataset Description . . . 156
7.3.2 Data Preprocessing . . . 157
7.3.2.1 Data Cleaning . . . 157
7.3.2.2 Data Normalization . . . 157
7.3.3 Experimental Framework . . . 158

7.3.3.1 Objective 1: Comparison of Spectral Bands and Existing Spectral Indices with ML Models 158
7.3.3.2 Objective 2: Development of a New Spectral Index for Soil pH and SOC 160
7.3.3.3 Multiple Feature Selection Techniques 162
7.4 Results 162
7.4.1 Objective 1: Assessment of Existing Spectral Indices 162
7.4.2 Objective 2: Screening Results of Multiple Feature Selection Techniques 164
7.4.3 Objective 2: Index-Based Analysis of Soil pH and SOC 166
7.5 Discussion 171
7.5.1 Advantages of Proposed Spectral Indices 171
7.5.2 Limitations of Proposed Spectral Indices 172
7.6 Summary 173

Chapter 8: Conclusion, Future Work & Social Impact **174**
8.1 Conclusion 174
8.1.1 Key Insights from This Research 176
8.1.2 Limitations 176
8.1.3 Future Work 177
8.2 Social Impact and Applications 178
8.2.1 Social Impact 179
8.2.2 Applications 179

References **184**

Publications proofs **223**

Plagiarism Report **238**

Curriculum Vitae **239**

Biography **240**

List of Tables

2.1 Spectral regions, wavelength ranges, and common remote sensing applications. 28
2.2 Advantages and limitations of spaceborne sensors [7] 36
2.3 Advantages and limitations of airborne sensors [7] 37
2.4 Advantages and limitations of UAVs [8] 38
2.5 Advantages and limitations of ground-based sensors [7] 39
2.6 Difference between multispectral and hyperspectral data [9, 10] 45

3.1 Open soil spectral datasets available for soil nutrient prediction 60
3.2 Mapping of open-access soil spectral datasets with research objectives 60
3.3 Summary statistics for LUCAS-2009 soil nutrients 61
3.4 Summary statistics for LUCAS-2015 soil nutrients 64
3.5 Summary statistics for AfSIS soil nutrients 66
3.6 Summary statistics of HYPERVIEW soil nutrients 68
3.7 Summary statistics of Rupnagar dataset soil nutrients 69
3.8 Summary statistics of phosphorus content in Univalle dataset 71
3.9 Previous studies using feature selection techniques for soil nutrients prediction 75
3.10 Previous studies on ML-based soil nutrients prediction 81
3.11 Previous studies on DL-based soil nutrients prediction 87
3.12 Different types of Vegetation indices used for soil nutrient prediction. . 93
3.13 Different types of soil and water indices used for soil nutrient prediction 94
3.14 Different types of temperature and thermal Indices used for soil nutrient prediction 94
3.15 Previous studies on soil nutrient prediction through spectral indices . . 96

4.1 Comparison of previous studies based on Lucas topsoil database and multispectral data for soil nutrients prediction 103
4.2 Selected hyperparameters of regression models and HyPLS-GPR . . . 111
4.3 Strengths and weaknesses of ML models and HyPLS-GPR 111
4.4 Comparison of ML models and HyPLS-GPR for soil nutrient prediction 114
4.5 Validation of HyPLS-GPR on Rupnagar dataset [5] 117

5.1 Previous work on HYPERVIEW dataset [4] 123
5.2 Hyperparameters outline the design of the CNViS-Net 126
5.3 Selected parameters of ML models combined with CNN and Transformer 126
5.4 Comparison of CNViS-NET for phosphorus (P) prediction 129
5.5 Comparison of CNViS-NET for potassium (K) prediction 129
5.6 Comparison of CNViS-NET for magnesium (Mg) prediction 130
5.7 Comparison of CNViS-NET for soil pH prediction 130
5.8 Validation of CNViS-Net on Univalle dataset [6] for soil phosphorus prediction 133

6.1 Comparison of CAT-MTLNet with SOTA methods for soil nutrient prediction 138
6.2 CAT-MTLNet architecture parameters and training configuration . . . 141
6.3 Results of CAT-MTLNet on testing set 145
6.4 R^2 Comparison of feature selection techniques with Transformer-MLP on testing set 146
6.5 Top 20 spectral bands selected using CAE for each soil nutrient prediction 147
6.6 Comparison with existing models 148

7.1 Previous studies on soil pH and SOC prediction through spectral Indices 154
7.2 Previous studies on soil pH and SOC prediction using AfSIS [3] and LUCAS-2009 [1] 155
7.3 Results of spectral bands and existing spectral indices used as input in ML models for soil pH prediction 163
7.4 Results of spectral bands and existing spectral indices used as input in ML models for SOC prediction 164

7.5 Selected hyperparameters of various ML models for spectral indices . 165
7.6 Results of various feature selection techniques with ANN for soil pH prediction 166
7.7 Results of various feature selection techniques with ANN for SOC prediction 166
7.8 Comparison of proposed SPI index with previous work on soil pH prediction using spectral bands 168
7.9 Comparison of proposed SOCI index with previous work on SOC prediction using spectral bands 169

List of Figures

1.1 Electromagnetic Spectrum 10
1.2 Automation workflow for spectral-based soil nutrient prediction 15

2.1 Active and passive remote sensing [11] 24
2.2 Spectral imaging showing three-dimensional spectral cube with spatial dimensions (x, y) and spectral dimension (z), each pixel holding a full spectral profile [12]. 33
2.3 Sensors used to capture spectral data [13] 34
2.4 Soil spectral data acquisition using remote sensing platforms [14] . . . 35
2.5 Multispectral data bands and the spectral response [15] 40
2.6 Hyperspectral data bands and the spectral response [15] 42
2.7 Comparison of the image stacks in multispectral and hyperspectral data. [16] 44
2.8 Automated workflow of preparing spectral data for ML models. 49
2.9 Satellite image preprocessing steps 49
2.10 Savitzky-Golay smoothing 52
2.11 Whittaker smoothing 53
2.12 Fourier Transform smoothing 53

3.1 Spatial distribution of LUCAS-2009 [1] 61
3.2 Spatial distribution of LUCAS-2015 [2] 63
3.3 Spatial distribution of AfSIS dataset [3] 65
3.4 (a) HYPERVIEW dataset [4] instance illustrating an image of the raw hyperspectral band (b) the corresponding masked region of interest. [17] 67
3.5 Spatial distribution of Rupnagar region [5] 69
3.6 Univalle database of reflectance spectra of 152 soil samples [18]. . . . 71

3.7 Usage of ML and DL models in previous studies for soil nutrient prediction 91

4.1 Research methodology flowchart for proposed HyPLS-GPR model . . 105
4.2 Top soil nutrients maps of Belgium using ArcMap/ArcGIS [19] 106
4.3 Image preprocessing steps of Landsat-8 multispectral images 107
4.4 Landsat-8 bands [20] 108
4.5 Architecture of proposed HyPLS-GPR model 113
4.6 RMSE and RPD bar plots representing the performance of HyPLS-GPR 116

5.1 Architecture of the proposed CNViS-Net framework. 125
5.2 Performance metrics bar plots of the proposed CNViS-Net framework 131
5.3 Scatter plots of the predicted values Vs. measured values of the proposed CNViS-Net framework 132

6.1 Architecture of the proposed CAT-MTLNet framework 139
6.2 Radar plots of RMSE showing the performance comparison of all feature selection techniques with Transformer-MLP (LUCAS-2009: (SOC, Ca) in $g.kg^{-1}$; P, in $mg.kg^{-1}$) 146

7.1 Research methodology flowchart for the development of new spectral indices 156
7.2 Flowchart for comparing spectral bands and existing spectral indices (VI- vegetation indices, SI- soil indices, WI- water indices, TTI- temperature and thermal indices) 159
7.3 Flowchart for Design and Validation of New Soil Spectral Indices . . . 159
7.4 Architecture to Compute SPI and SOCI Indices using PCA and ANN . 160
7.5 Proposed SPI and SOCI indices training & testing error for soil pH and SOC prediction 170
7.6 Proposed SPI and SOCI indices residual plot for soil pH and SOC prediction 171

8.1 Climate-smart land use for sustainable soil and carbon management. . 180
8.2 AI-driven precision farming using drones and sensors [21]. 181

8.3 Digital platforms empower data-driven and smart soil nutrient management. 181
8.4 UAV-assisted spectral monitoring for soil nutrient prediction. 182
8.5 Data-driven soil nutrient models guiding agricultural policy and subsidies. 183

List of Abbreviations

AfSIS Africa Soil Information Service 65
AI Artificial Intelligence 8
AFGP Attention-Guided Feature Pruning 137
CNN Convolutional Neural Network 18
CSV Comma Separated Value 157
CAE Concrete Autoencoder 21
DL Deep Learning 8
DN Digital Number 29
DCNN Deep Convolutional Neural Network 103
DRS Diffuse Reflectance Spectroscopy 149
DOS Dark Object Subtraction 39
ELM Extreme Learning Machine (ELM) 91
EMS Electromagnetic Spectrum 9
EMR Electromagnetic Radiation 23
ELM Empirical Line Method 91
FFT Fast Fourier Transform 53
GPR Gaussian Process Regression 18
GPS Global Positioning System 37
IMU Inertial Measurement Unit 37
KNN K-Nearest Neighbors 127
LN Layer Normalization 125
LUCAS Land Use and Coverage Area Frame Survey 60

LiDAR Light Detection and Ranging 24
MHSA Multi-Head Self-Attention 125
ML Machine Learning 8
MLP Multi-Layer Perceptron 21
NMSA National Mission on Sustainable Agriculture 183
NDVI Normalized Difference Vegetation Index 95
PCA Principal Component Analysis 15
PCC Pearson Correlation Coefficient 15
PLSR Partial Least Square Regression 18
RFE Recursive Feature Elimination 15
RL Reinforcement Learning 137
RNN Recurrent neural network 90
RKVY Rashtriya Krishi Vikas Yojana 183
SAR Synthetic Aperture Radar 24
SDG Sustainable Development Goals 1
SOCI Soil Organic Carbon Index 18
SPI Soil pH Index 18
SPA Successive Projections Algorithm 15
SVR Support Vector Regression 21
SNV Standard Normal Variate 14
ViTs Vision Transformers 18

Chapter 1

Introduction

Agriculture is fundamental in ensuring food security, supporting livelihoods, and sustaining rural economies [22]. The United Nations Sustainable Development Goals (SDG) 2—Zero Hunger—underscores the urgent need to enhance agricultural productivity and promote sustainable food production systems by 2030 [23]. The pressure on agricultural practices to produce more food using fewer resources intensifies due to the continuous rise in the global population, with billions of mouths to feed [24]. This challenge demands increased production and innovative, sustainable, and environmentally responsible practices.

At the heart of this transformation lies the soil—a vital, yet often overlooked, component of the global food system [25]. Soil is not merely a medium for plant growth; it is a complex ecosystem that supplies water, anchors roots, and delivers essential nutrients required for healthy crop development [26]. Among its many roles, the nutrient-supplying function of soil is perhaps the most critical for determining crop yield, quality, and resilience.

Soil analysis becomes indispensable to meet future food demands. Farmers and agricultural stakeholders can make informed decisions directly influencing crop productivity by assessing soil characteristics, especially its nutrient content and health indicators [27]. Based on the quantity required and the biological function, soil nutrients are broadly categorized as [28]:

1. ***Macronutrients:*** These are essential elements that plants need in large amounts to support key physiological functions which include photosynthesis, protein

synthesis, and enzyme activation. Nitrogen (N), Phosphorus (P), and Potassium (K) are primary macronutrients, while Sulfur (S), Calcium (Ca), and Magnesium (Mg) are secondary macronutrients crucial for structural development and metabolic functions [29].

2. ***Micronutrients:*** These are required in much smaller quantities than macronutrients, are equally vital for plant growth and development. Elements like Chlorine (Cl), Iron (Fe), Boron (B), Manganese (Mn), Zinc (Zn), Copper (Cu), and Molybdenum (Mo) play key roles in enzymatic reactions, hormone regulation, and overall plant immunity and productivity [30].

3. ***Other soil indicators:*** The growth of crops relies on several other crucial soil indicators, which are listed below [31]:-

   (a) ***Soil pH:*** It is a critical factor that influences nutrient availability, microbial activity, and overall soil health [32]. Values outside the optimal range of 5.5 to 7.5 can lead to nutrient deficiencies or toxicities, hindering plant growth and productivity. Acidic soils (pH $<$ 5.5) may reduce the availability of essential nutrients like phosphorus, while alkaline soils (pH $>$ 7.5) can limit the uptake of micronutrients such as iron and zinc. Therefore, maintaining an optimal pH is essential for balanced nutrient uptake and healthy crop yield [33].

   (b) ***Soil Organic Carbon (SOC):*** It exclusively refers to the carbon atom in organic molecules. It is essential for synthesizing organic acids in soil, which are necessary for dissolving soil minerals, increasing plant availability, and leaching nutrients. Additionally, SOC improves soil structure, enhances water retention, and supports microbial activity, contributing to long-term soil fertility and sustainable plant growth [34].

   (c) ***Soil Organic Matter (SOM):*** It is primarily made up of carbon, hydrogen, and oxygen. At the same time, organic wastes also contain trace amounts of nitrogen, phosphorus, sulfur, potassium, calcium, and magnesium. SOM enhances soil structure by increasing aggregation and reducing soil compaction. It enhances water infiltration and aeration, which are essential

for root growth and nutrient uptake [35].

Plants require a proper balance of nutrients to maintain soil fertility and promote healthy crop growth. An optimal nutrient balance supports key physiological functions such as root development, flowering, and resistance to pests and diseases, ultimately enhancing crop quality and productivity [36]. Conventional farming practices often rely on the uniform application of fertilizers to balance nutrients. However, this approach overlooks site-specific nutrient requirements, resulting in inefficient resource utilization, economic losses, and environmental degradation [37].

Hence, understanding the existing nutrient profile of the soil is crucial for implementing precision agriculture strategies. It enables the identification of nutrient deficiencies or toxicities and supports decisions on fertilizer application, crop selection, and land management [38]. Such predictive insights help minimize over-fertilization, reduce input costs, and mitigate the ecological impacts of agriculture. In this way, accurate and timely soil nutrient prediction enhances crop yields and contributes to broader sustainability goals [39, 40].

1.1 Methods to Predict Soil Nutrients

1.1.1 Laboratory-based Chemical Analysis:

This traditional method involves chemical testing procedures such as wet chemistry and titration to assess nutrient concentrations in soil samples [41]. These highly accurate techniques serve as a benchmark for validating newer methods. Several studies have utilized laboratory-analyzed soil chemical data as input for ML models to predict soil nutrients [42–45]. These were aimed at identifying correlations among soil parameters that enable the prediction of additional properties without direct measurement.

Farwa et al. [42] used lab-measured macronutrient concentrations to train models for predicting nitrogen, phosphorus, and potassium levels. In contrast, Sunori et al. [43] relied on chemical pH measurements to develop ML-based soil pH prediction models [43]. Similarly, Mahmoudzadeh et al. [44] utilized laboratory-determined soil organic matter and organic carbon values to calibrate models that mapped these properties across croplands. Wilhelm et al. [45] incorporated chemical soil health

indicators alongside microbiome data for supervised ML prediction of soil health. These studies highlight the accuracy and reliability of chemical lab data, which serves as a gold standard for ML-based prediction.

Advantages: The laboratory methods offer unmatched accuracy, reproducibility, and comprehensiveness, especially for complex or subtle nutrient profiles [45]. These are essential for benchmarking emerging technologies, as well as training and validating predictive models across diverse soil types and geographic regions.

Disadvantages: Despite the accuracy, laboratory-based chemical analyses come with several significant limitations that hinder the utility for large-scale, real-time, and cost-effective soil monitoring in modern agriculture, which are as follows [31, 46, 47]:

1. ***Time-Consuming Process:*** Traditional soil analysis involves multiple steps, including sample collection, drying, grinding, chemical extraction, and laboratory testing—all of which are time-intensive. This prolonged duration makes it impractical for scenarios requiring rapid decision-making, such as precision agriculture interventions during a cropping season [48].

2. ***High Cost Per Sample:*** These methods require specialized chemicals, high-precision instruments, and controlled laboratory environments, resulting in substantial operational costs. Additionally, recurring expenses such as transportation, technician wages, and lab maintenance further escalate the cost per sample, making the approach unaffordable for resource-constrained regions or smallholder farmers [49].

3. ***Labor-Intensive Operations:*** Human involvement is required at nearly every stage—from field collection to chemical testing and result interpretation. This dependence on skilled personnel introduces the risk of human error and creates logistical bottlenecks when scaling up soil testing efforts across multiple farms or regions [50].

4. ***Limited Scalability:*** Centralized laboratory infrastructure cannot handle high-throughput testing over large agricultural landscapes. The physical limitations of lab capacity, staff availability, and turnaround times severely restrict the ability to monitor soil health at a regional or national scale [51].

5. ***Lack of Temporal Responsiveness:*** Since traditional methods are not real-time, these approaches fail to capture the dynamic changes in soil nutrient status that occur due to rainfall, irrigation events, fertilization, or plant uptake. This limitation makes it difficult to monitor seasonal or even intra-seasonal variations in soil health, which are crucial for adaptive nutrient management [52].

1.1.2 Imaging-based soil nutrient prediction:

In recent years, imaging-based techniques have emerged as powerful alternatives to traditional laboratory methods for assessing soil nutrients. These approaches rely on capturing visual or spectral information from soil samples or fields and using computational models to infer nutrient concentrations. In the electromagnetic spectrum, RGB bands capture only the visible range, while spectral sensors extend into the Near-Infrared (NIR), Short-wave Infrared (SWIR), and mid-infrared spectral regions, which are highly sensitive to soil chemical and physical properties. These broader spectral regions provide the foundation for machine learning–based soil analysis, as discussed in detail in chapter 2. Imaging-based methods are broadly categorised as follows:

1.1.2.1 RGB Imaging:

It leverages standard red, green, and blue bands to produce ”true colour” visuals similar to human vision, covering bands in the electromagnetic spectrum in the range of 400–700 nm [53,54]. RGB cameras—available in smartphones, drones, and consumer-grade digital devices—enable the acquisition of soil images in diverse conditions and at large scales. Its simplicity and accessibility make it a suitable early candidate for visual soil assessment. When paired with machine learning, these images can extract meaningful patterns from soil appearance, offering a faster and lower-cost alternative to traditional laboratory-based nutrient prediction methods, especially when resources are limited [55].

Several studies have highlighted the potential of RGB imaging combined with machine learning for predicting soil nutrients by capturing variations in soil color and texture that correlate with chemical attributes [56–60]. Barman et al. [56] used smartphone images and supervised learning to estimate soil pH. In contrast, Natarajan

et al. [57] applied gradient-boosted regression on RGB images from remote sensing, confirming the utility of RGB even without extended spectral data. However, these methods can be affected by environmental factors, such as lighting and moisture.

To address such variability in RGB-based predictions, Gorthi et al. [58] used smartphone-captured RGB images along with reflectance normalization and image perturbation techniques to improve the prediction of soil organic matter under uncontrolled lighting. Similarly, Fu et al. [59] leveraged RGB imagery to estimate SOM under varying soil moisture conditions, highlighting the importance of preprocessing to mitigate reflectance distortions and enhance model robustness. More recently, Lavanya et al. [60] showed that a smartphone-based RGB device, combined with the ascorbic acid extraction method, can accurately predict available phosphorus. This study employed an RF and digital soil mapping model with synthetic data augmentation, achieving R^2 = 0.73, which outperformed the lab method (R^2 = 0.69), confirming the field-level viability of low-cost RGB tools for nutrient mapping.

Advantages: RGB imaging offers several practical advantages for soil analysis. It is highly cost-effective, as RGB cameras are widely available in consumer devices, such as smartphones and drones, eliminating the need for specialized or calibrated hardware. The approach offers operational simplicity, enabling users to directly capture and process soil images in the field without requiring specialized laboratory infrastructure. Furthermore, the integration of machine learning enables real-time analysis, where soil nutrients can be estimated quickly and efficiently to support precision agriculture decision-making.

Disadvantages: Despite its widespread use for visual interpretation, RGB data suffers from several critical limitations that restrict its applicability in scientific remote sensing analysis, such as [53]:

1. ***Limited Spectral Range:*** RGB sensors operate only within the visible spectrum (approximately 400–700 nm). These sensors do not capture radiation in the NIR, SWIR, and TIR regions, which are critical for detecting vegetation stress, soil moisture, and material differences.

2. ***Low Spectral Resolution:*** RGB systems collect data in three broad bands, meaning many subtle spectral features are missed or merged. Materials with

similar visible color but different chemical properties appear indistinguishable. No unique spectral signature can be derived from RGB images [61].

3. ***Not Suitable for Quantitative Analysis:*** The lack of calibration for precise scientific measurements renders RGB images primarily qualitative. It cannot be used directly for tasks like vegetation index calculation, soil nutrient prediction, or mineral mapping [54].

4. ***Insensitive to Non-Visible Variability:*** Many surface features show greater NIR and SWIR bands variation than visible bands. RGB images fail to detect early plant stress, water stress, or soil composition differences otherwise visible in multispectral or hyperspectral data [53].

5. ***Affected by Atmospheric and Illumination Conditions:*** Since RGB relies on visible light, it is more prone to illumination changes, shadows, haze, and cloud cover, which can distort true color interpretation [54].

1.1.2.2 Spectral Imaging:

While RGB images are visually intuitive, this format lacks the spectral detail needed for precise scientific analysis [36]. This limitation underscores the value of remote sensing-based spectral imaging, which utilizes a significantly broader range of the electromagnetic spectrum. It includes the visible, near-infrared, and short-wave infrared regions, enabling accurate material classification and detailed surface analysis [62]. The visible bands capture color and texture information, while the near infrared region is susceptible to soil organic matter, moisture content, and certain nutrient-related absorption features [63]. Short-wave infrared bands provide additional sensitivity to mineralogical and chemical compositions, enabling more accurate discrimination of soil types and nutrient levels [64]. It includes multispectral and hyperspectral data, which emerged as a powerful approach for soil nutrient assessment [65–67] (described in Chapter 2).

Disadvantages: Spectral imaging requires careful preprocessing to remove noise, correct for atmospheric and illumination effects, and calibrate sensor responses. Environmental factors, such as moisture, surface roughness, and shadows, can introduce

artifacts that affect the accuracy of nutrient prediction [67].

Advantages: Despite the challenges, remote sensing–based spectral imaging provides a non-destructive, scalable, and high-precision approach for monitoring soil nutrients across large areas. It makes it a crucial advancement over conventional and RGB-based methods. The advantages of spectral imaging are as follows:

1. ***Captures Detailed Soil Variability:*** Spectral imagery detects subtle variations in soil composition that are often missed by point-based laboratory sampling or RGB imagery [65]. It enables a more precise understanding of spatial heterogeneity, critical for effective soil management.

2. ***Non-invasive and Rapid Analysis:***: Unlike laboratory tests, which are time-consuming and destructive, spectral data provides quick, non-destructive assessments of soil nutrient levels. It allows repeated measurements over time without disturbing the soil structure [41].

3. ***Spatially Continuous Coverage:*** It overcomes the spatial limitations of laboratory sampling by providing continuous data over large agricultural areas, unlike RGB images, which only provide visual information for limited points. This coverage ensures that variations across entire fields are captured, improving the reliability of soil assessments [68].

4. ***Enhanced Spectral Information:*** Spectral imagery captures wavelengths beyond visible light (VIS, NIR, SWIR), providing more accurate insights into soil chemical and physical properties than RGB imagery [66]. This richer spectral information helps identify invisible nutrient deficiencies that are not visible to the human eye [69].

5. ***Facilitates data-driven prediction:*** Spectral imagery can be integrated with machine learning models to predict soil nutrients across fields, overcoming the limited predictive capability of conventional laboratory or visual methods [70]. It enables scalable, data-driven decision-making for precision agriculture.

Hence, this thesis aims to integrate remote sensing data with Artificial Intelligence (AI) approaches, including Machine Learning (ML) and Deep Learning (DL) for robust

soil nutrient prediction. It can help to overcome the limitations of conventional laboratory testing and RGB imaging methods by enabling continuous, large-scale assessment of nutrient status. Such integration can facilitate optimized fertilizer usage, enhanced monitoring of soil health, and data-driven support for sustainable precision agriculture.

1.2 Motivation and Background

1.2.1 Why Remote Sensing for Soil Nutrient Prediction?

Remote sensing is a powerful, non-contact method of acquiring information about the Earth's surface by detecting the electromagnetic radiation that is reflected or emitted from it [71]. It plays a pivotal role in soil science by enabling large-scale and continuous monitoring of soil nutrients [72]. In soil nutrient prediction, remote sensing facilitates the analysis of spatial and temporal variability in soil characteristics, thereby supporting efficient and site-specific agricultural management [73]. Unlike point-based laboratory testing, it enables scalable soil assessment, especially when integrated with spectral data and ML models.

1.2.2 Electromagnetic Spectrum

The Electromagnetic Spectrum (EMS) represents the complete range of electromagnetic radiation, which consists of oscillating electric and magnetic fields that travel through space at the speed of light [74]. This radiation is characterized by its wavelength (λ) or frequency (ν), and different materials interact with other parts of the spectrum in unique ways. Like other materials, soil reflects, absorbs, and emits electromagnetic energy differently across this spectrum. Electromagnetic radiation refers to energy transmitted through waves or particles (photons) [75]. Equation 1.1 shows the relationship between wavelength, frequency, and the speed of light (c).

$$c = \lambda \cdot \nu \tag{1.1}$$

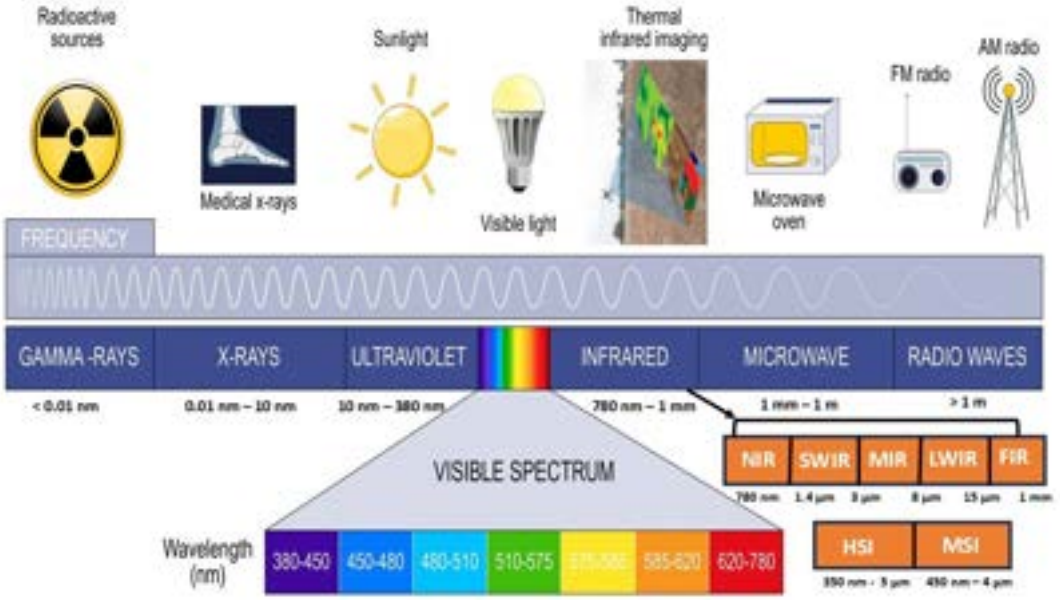

Figure 1.1: Electromagnetic Spectrum

where, c = speed of light ($\approx 3 \times 10^8$ m/s), λ = wavelength (in meters), ν = frequency (in Hz).

Figure 1.1 illustrates the wavelength divisions fundamental to understanding how materials interact with electromagnetic energy. These divisions illustrate the continuous nature of the electromagnetic spectrum, spanning from shorter wavelengths, such as gamma rays, to longer wavelengths, like radio waves. This categorization also forms the foundation for designing sensors, communication systems, and other electromagnetic-based technologies.

1.2.2.1 Electromagnetic Spectrum Components and Applications

The electromagnetic spectrum encompasses a range of wavelengths and frequencies, each with distinct properties and applications across various scientific and technological domains, such as [76]:

1. ***Gamma Rays:*** These are the most energetic form of electromagnetic radiation with extremely short wavelengths. The wavelength is less than 0.01 nanometers (nm). The applications include medical treatments (e.g., cancer radiotherapy), sterilization of equipment, and nuclear research.

2. ***X-Rays:*** These are high-energy radiation capable of penetrating most substances, used mainly in imaging. The wavelength ranges from 0.01 to 10 nm. The applications include medical imaging, security scanning, and crystallography.

3. ***Ultraviolet (UV) Rays:*** These rays extend beyond the visible spectrum and possess higher energy than visible light. The wavelength ranges from 10 to 380 nm. The applications include disinfection, forensic analysis, and the study of plant health and mineral fluorescence.

4. ***Visible (VIS) Light:*** This is the portion of the spectrum visible to the human eye, spanning all colors. The wavelength ranges from 380 to 780 nm. The applications include human vision, optical imaging in remote sensing, and photography. The color bands are as follows: violet (380–450 nm), blue (450–480 nm), cyan (480–510 nm), green (510–575 nm), yellow (575–585 nm), orange (585–620 nm), and red (620–780 nm).

5. ***Infrared (IR) Rays:*** These are just beyond the red edge of the visible spectrum and are perceived as heat. The wavelength ranges from 780 nm to 1 mm. The applications include thermal imaging, environmental monitoring, soil and vegetation analysis, and night vision systems. The subregions are as follows: Near-Infrared (NIR): 780–1400 nm, Short-wave Infrared (SWIR): 1400–3000 nm, Mid-Infrared (MIR): 3–8 μm, and Thermal Infrared (TIR): 8–15 μm.

6. ***Microwaves:*** These are longer-wavelength radiation, which is used for atmospheric sensing and communication. The wavelength ranges from 1 mm to 1 meter. The applications include weather radar, satellite communication, remote sensing (e.g., Synthetic Aperture Radar or SAR), and GPS.

7. ***Radio Waves:*** This is the most extended wavelength portion of the EM spectrum, widely used in wireless transmission. The wavelength ranges from greater than 1 meter. The applications include radio broadcasting, mobile communication, television signals, and astronomy.

The most relevant spectral regions for soil analysis include the VIS, NIR, SWIR, and MIR regions [72]. These regions capture essential soil features such as color, organic

matter, moisture, and mineral composition, making them crucial for remote sensing-based nutrient prediction.

1.2.2.2 Key Concepts in Remote Sensing for Soil Nutrient Prediction

In remote sensing applications for soil nutrient prediction, several foundational concepts—such as spectral bands, spectral reflectance, spectral signatures, and spectral indices are used to interpret how soil and other surface materials interact with electromagnetic radiation [73]. These concepts help extract meaningful information about soil composition based on its response to different wavelengths of light.

1. ***Spectral Bands:*** It refers to specific, contiguous wavelength intervals within the electromagnetic spectrum where remote sensing sensors record reflectance or radiance data [69]. Each band captures unique information related to the physical and chemical properties of surface materials, enabling a detailed analysis of soil characteristics.

2. ***Spectral Reflectance:*** It refers to the proportion of incident electromagnetic radiation that is reflected by a surface as a function of wavelength [77]. Spectral reflectance provides critical insights into soil composition, structure, and condition, making it a key parameter in soil analysis.

3. ***Spectral Signature:*** It is the unique pattern of reflectance exhibited by a material over a range of wavelengths [78]. Each material reflects, absorbs, and transmits electromagnetic radiation differently, resulting in distinctive spectral curves when plotted as reflectance against wavelength.

4. ***Spectral Indices:*** These are mathematical combinations of reflectance values from specific bands designed to highlight particular soil or vegetation properties [79]. In soil analysis, moisture, organic matter, and bare soil exposure indices are particularly useful [30]. However, a key gap is the limited availability of novel spectral indices designed explicitly for soil nutrient prediction.

1.2.3 Spectral Data and Spectral Imaging

Spectral data and imaging enable the observation and analysis of materials, such as soil, by capturing how light reflects at different wavelengths, thereby helping us understand the physical and chemical properties without direct contact.

1. ***Spectral Data:*** It provides a non-invasive, scalable, and efficient means of analyzing the Earth's surface by measuring how different materials interact with electromagnetic radiation [80]. Spectral data can be captured in multiple forms, such as direct reflectance measurements from a single point, area, or images through spectral imaging.

2. ***Spectral Imaging:*** It records image data across multiple, narrow wavelength bands of the electromagnetic spectrum, allowing detailed characterization of surface materials [12]. Spectral images generate a three-dimensional data cube, where two dimensions represent spatial information, and the third dimension captures spectral information across multiple wavelengths.

1.2.4 Types of Spectral Data: Multispectral and Hyperspectral

Spectral data primarily fall into two categories—multispectral and hyperspectral—each differing in the number, width, and continuity of spectral bands captured. A more detailed discussion is provided in chapter 2. The spectral data types are as follows:

1. ***Multispectral Data (MSD):*** It captures the data at a few selected and discrete spectral bands across the electromagnetic spectrum [81]. This technique provides valuable information about the Earth's surface by measuring reflected light in several broad wavelength regions.

2. ***Hyperspectral Data (HSD):*** It is an advanced data type that captures the full reflectance spectrum [82]. Unlike MSD, which samples a few broad spectral bands, HSD measures hundreds of narrow, contiguous spectral bands. It provides a continuous spectral profile that can uniquely characterize different materials based on the spectral signatures.

1.3 Automated Spectral-based Machine Learning for Soil Nutrient Prediction

Automating spectral data is crucial to ensure clean, consistent, and model-ready input for ML and DL models. Figure 1.2 illustrates an automated workflow for predicting soil nutrients using spectral data and AI models, enabling rapid, non-destructive, and scalable analysis in precision agriculture. The process begins with soil sampling and laboratory analysis to obtain ground-truth data, followed by the acquisition and preprocessing of spectral images. The spectral data matrix is refined through noise removal, normalization, and feature selection before being used in AI models. Finally, soil nutrient predictions are evaluated using performance metrics such as R^2, RMSE, and MAE, as discussed in chapter 3. The details of the automation workflow are as follows:-

1. **Conversion of Raw Spectral Images into Spectral Reflectance Values:** The process begins by converting raw spectral images into structured spectral datasets. It includes band composition, mosaicing, clipping, and spatial selection. These steps transform unprocessed sensor data into spatially aligned and analyzable formats linked with ground truth information [5].

2. **Noise Removal:** Spectral data often contains noise due to atmospheric interference, sensor limitations, or environmental factors [83]. This thesis employs noise-reduction techniques, including Savitzky-Golay smoothing, Whittaker filtering, and the Fourier Transform, to preserve meaningful spectral features.

3. **Spectral Normalization:** This thesis employs normalization techniques, including Standard Normal Variate (SNV) and the Standard Scaler, to standardize the spectral reflectance values across all samples. These methods address external inconsistencies, such as surface roughness or illumination variation, which can negatively affect model accuracy [84].

4. **Feature Selection:** It helps in reducing data dimensionality while retaining informative features, which is critical for effective modeling [85]. Techniques

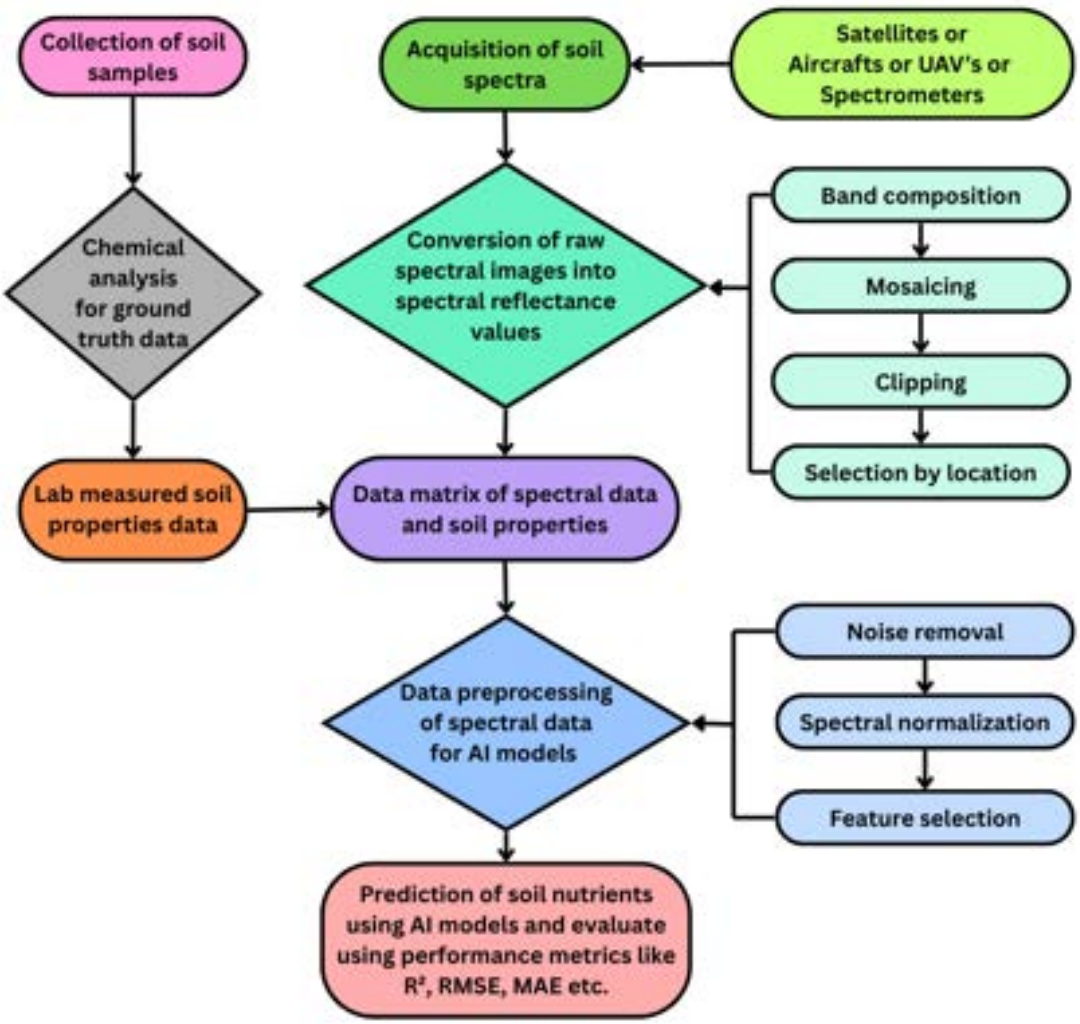

Figure 1.2: Automation workflow for spectral-based soil nutrient prediction

such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Successive Projections Algorithm (SPA), Pearson Correlation Coefficient (PCC), LASSO, and XGBoost feature importance are commonly used to eliminate redundant or irrelevant features. It enhances model performance, reduces overfitting, and improves interpretability.

1.3.1 Research Gaps

1. **Limited coverage of DL models:** Despite the ability to capture complex, non-linear relationships in high-dimensional hyperspectral data, the application of DL models for soil nutrient prediction remains limited. Most existing studies mainly rely on ML methods, which may lack the advanced feature extraction capabilities

of DL [86–89]. While ML approaches perform reasonably well, DL models offer greater potential to exploit rich spectral information, yet remain underutilized in this domain.

2. **Limited research on selecting the optimal spectral band subset for soil nutrient prediction using hyperspectral data:** Feature selection is a crucial step in hyperspectral data analysis for soil nutrient prediction, as it helps reduce data dimensionality and enhances model performance [85]. However, comprehensive research is still lacking on identifying the optimal subset of spectral bands that capture the most relevant information for nutrient prediction [90–93]. Determining these bands is crucial for improving model efficiency and scalability.

3. **Underexplored development of novel spectral indices:** Although spectral indices have been widely used for vegetation monitoring, the development of tailored indices for soil nutrient prediction remains underexplored [30, 50, 79]. A recent study (Gulhane et al. [30]) used multispectral data and proposed novel indices for soil pH, iron, and phosphorus, highlighting the potential. However, similar advancements using hyperspectral data are still lacking, leaving a notable gap in current research.

4. **There is a lack of studies using existing spectral indices for soil nutrient prediction with hyperspectral datasets:** Despite the potential of spectral indices in remote sensing, there is a notable lack of studies that apply existing spectral indices (vegetation, soil, water, temperature, and thermal) for predicting soil nutrients using hyperspectral datasets. In contrast, multispectral datasets have been widely used for such applications, demonstrating the utility of spectral indices in simpler settings. However, existing hyperspectral research often overlooks the potential benefits of leveraging these indices, which could simplify the feature selection process and improve prediction performance. Investigating the effectiveness of spectral indices in predicting soil nutrients from hyperspectral data could lead to more efficient and interpretable models [30, 50].

1.4 Problem Statement

This thesis addresses the critical challenge of optimizing soil nutrient prediction using spectral data for precision agriculture through robust ML and DL frameworks. In particular, (1) it targets the problem of high dimensionality inherent in hyperspectral datasets, which leads to increased computational complexity and reduced model efficiency using ML and DL models. (2) This thesis focuses on selecting efficient, informative spectral bands to enhance model interpretability and scalability, beyond just accuracy. (3) It integrates existing spectral indices—an underexplored area in current literature—alongside hyperspectral data for nutrient prediction. (4) In addition, this thesis emphasises the development of novel spectral indices specifically tailored for soil nutrient prediction to improve the accuracy and robustness of the model.

1.5 Research Objectives and Contribution of the Thesis

The main objective of this thesis is the efficient prediction of soil nutrients using spectral data and ML techniques. It achieves the following objectives:

- **Objective 1: Critical Review -** To perform a critical review for soil nutrient prediction using artificial intelligence and spectral data.

  Contribution: This thesis conducts a critical, systematic literature review to examine the strengths and weaknesses of various ML and DL methods used for predictive models for soil nutrient prediction. The literature review examines various feature selection techniques and the application of spectral indices, along with a review of open spectral soil datasets. It also highlights the open challenges in existing literature, including the lack of novel spectral indices and hybrid ML/DL frameworks. (Completed by **Journal 1**).

- **Objective 2: Feature Selection -** To compare existing feature selection techniques and propose the most suitable method for identifying soil nutrients.

  Contribution: This thesis evaluates multiple feature selection techniques for soil nutrient prediction using hyperspectral datasets (LUCAS-2009 [1] and AfSIS [3]). It compares the effectiveness in selecting relevant spectral bands

and assesses the impact on model performance. The results indicate that PCA is the most suitable technique for achieving a balance between accuracy and computational efficiency. (Completed by **Journal 2**).

- **Objective 3: Deep Learning -** To compare ML/DL models for improved performance in soil nutrient prediction.

  Contribution: This thesis proposes a novel hybrid ML model and two DL-based frameworks for soil nutrient prediction using multispectral and hyperspectral data. The following points describe these novel frameworks:

  1) The proposed hybrid ML model *HyPLS-GPR* for multispectral soil nutrient prediction combines Partial Least Square Regression (PLSR) with Gaussian Process Regression (GPR). This hybrid model successfully forecasts multiple soil nutrients using multispectral Landsat-8 imagery and open soil datasets (LUCAS-2015 [2] and Rupnagar dataset [5]), expanding its utility for agricultural and environmental monitoring. (Completed by **Conference 2**).

  2) The proposed DL-based *CNViS-Net* framework introduces a self-attention mechanism for hyperspectral data analysis using HYPERVIEW [4] and Univalle [6] datasets, integrating Convolutional Neural Network (CNN) and Vision Transformers (ViTs) for robust feature extraction, followed by ML-based prediction using SVR. It demonstrates superior performance in capturing complex spectral patterns and achieves state-of-the-art accuracy for nutrient prediction. (Completed by **Conference 4**).

  3) The proposed *CAT-MTLNet*, a novel deep multi-target framework for hyperspectral soil nutrient prediction using LUCAS-2009 [1], and AfSIS [3] datasets. It integrates a Concrete Autoencoder (CAE) for data-driven spectral feature selection and a Transformer-MLP architecture for capturing long-range spectral dependencies, enabling joint prediction of multiple soil nutrients. (Completed by **Journal 3**).

- **Objective 4: Spectral Indices -** To explore, analyze, and evolve new spectral indices for soil nutrient prediction.

  Contribution: This thesis develops two novel spectral indices designed explicitly for the prediction of Soil Organic Carbon Index (SOCI) and Soil pH Index (SPI),

leveraging unique spectral signatures to enhance the accuracy and reliability of nutrient prediction. In addition to proposing these new indices, this work investigates the predictive potential of existing spectral indices, including vegetation, soil, water, temperature, and thermal indices, for SOC and soil pH prediction using hyperspectral datasets (LUCAS-2009 [1] and AfSIS [3]). (Completed by **Journal 2**).

1.6 Significance of Thesis

This thesis addresses several critical gaps in spectral-based soil nutrient prediction by proposing an integrated approach that combines optimal feature selection, advanced ML and DL frameworks, and novel spectral indices. It emphasizes model interpretability and generalizability through efficient spectral band selection. Including a detailed comparative analysis of multiple feature selection techniques, leading to the identification of PCA as a balanced solution, enhances both computational efficiency and prediction robustness. This targeted approach reduces redundancy in hyperspectral data, minimizes the risk of overfitting, and contributes to the development of interpretable, resource-efficient models suitable for large-scale deployment in precision agriculture.

Furthermore, this thesis contributes to the advancement of soil nutrient prediction by introducing three novel prediction approaches tailored to different data sources and modeling requirements. The hybrid ML model (HyPLS-GPR) proposed for multispectral data provides an accessible and scalable solution that utilizes satellite imagery, thereby expanding the applicability of soil nutrient prediction to resource-constrained settings. The CNViS-Net framework effectively captures the complexity of hyperspectral data using deep learning and attention-based architectures, achieving state-of-the-art accuracy in nutrient prediction. CAT-MTLNet, a novel deep multi-target framework for hyperspectral soil nutrient prediction, efficiently reduces dimensionality while maintaining high prediction accuracy, confirming its robustness for multi-target hyperspectral soil nutrient prediction. In addition, the development of two novel spectral indices—SOCI and SPI—and the thorough evaluation of existing indices establish a strong foundation for spectral index-based modeling. These indices

enhance model performance while reducing reliance on raw spectral bands, promoting generalization across varying soil and environmental conditions. Collectively, this thesis delivers a comprehensive, multi-dimensional contribution to remote sensing-based soil analysis by bridging the gaps between feature selection, model design, and index formulation.

1.7 Thesis Overview

The following shows the structure of the thesis, providing an overview of the content and purpose of each chapter:

Chapter 2: Technical Background of Remote Sensing for Soil Nutrient Prediction: This chapter introduces the theoretical foundation of spectral data and its relevance to remote sensing. It explains the electromagnetic spectrum, principles of spectral reflectance, and various imaging techniques used in remote sensing. The chapter discusses the key concepts in detail, including spectral bands, signatures, and indices, along with a brief overview of the calibration and correction procedures necessary for preprocessing spectral data. It provides the scientific basis necessary to understand and implement the models presented in later chapters.

Chapter 3: Literature Review: This chapter reviews advanced spectral-based soil nutrient prediction methods, highlighting machine and deep learning models, feature selection techniques, spectral indices, and prior research on soil analysis. It presents publicly available soil datasets used for spectral-based nutrient prediction, summarizing the spectral, chemical, and physical characteristics, including wavelength ranges, resolutions, sample counts, and measured nutrient contents. This review identifies research gaps, highlights methodological strengths and weaknesses, and motivates the novel contributions of this thesis.

Chapter 4: HyPLS-GPR: A Hybrid Machine Learning Model for Soil Nutrient Prediction Using Multispectral Data: This chapter proposes a hybrid machine learning framework using PLSR + GPR for multispectral-based soil nutrient prediction. It investigates how ML algorithms handle multispectral features with limited spectral resolution. The chapter highlights the model's ability to extract relevant soil information and suitability for operational agricultural monitoring. It further presents

the comparisons of different ML models (PLSR, Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), and GPR) to determine the effectiveness and robustness in real-world soil analysis tasks.

Chapter 5: CNViS-Net: A Self-Attention Framework for Soil Nutrient Prediction Using Hyperspectral Data: This chapter presents CNViS-Net, a novel self-attention-based framework for predicting soil nutrients using hyperspectral data. It effectively integrates an attention mechanism to model complex, non-linear relationships between soil and spectral properties. The chapter emphasizes the strength of DL architectures such as CNNs and Transformers. It proposes a DL-based framework in learning non-linear soil–spectral relationships and capturing fine-grained spectral details. The chapter further discusses the advantages of DL in addressing complex feature interactions and its potential for large-scale soil monitoring.

Chapter 6: CAT-MTLNet: A Novel Deep Multi-Target Framework for Hyperspectral Soil Nutrient Prediction Using Concrete Autoencoder and Transformer: This chapter introduces CAT-MTLNet, a novel deep learning framework for estimating multiple soil nutrients using hyperspectral data. The model leverages the representational power of deep architectures by integrating a Concrete Autoencoder (CAE) for data-driven spectral feature selection and a Transformer-based network for learning long-range spectral dependencies. The chapter details the deep multi-task learning design that allows the network to jointly predict multiple soil nutrients while sharing spectral knowledge across tasks. Comprehensive experiments show that CAT-MTLNet outperforms existing deep models, efficiently extracting key spectral features for accurate multi-target soil prediction.

Chapter 7: Development of Novel Spectral Indices for Soil pH and Organic Carbon Prediction: This chapter introduces the development of novel spectral indices designed explicitly for soil nutrient prediction, namely the Soil pH Index (SPI) and the Soil Organic Carbon Index (SOCI). It systematically describes the methodology for constructing these indices and selecting the most responsive spectral regions. The chapter evaluates the predictive performance against well-established indices using statistical and machine learning models. Comparative analyses demonstrate the added value of custom-designed indices tailored to soil-specific properties, offering an accessible yet powerful tool for practical agricultural monitoring and nutrient

management.

Chapter 8: Conclusion, Future Work & Social Impact: The final chapter summarizes the main findings and contributions of this thesis. It outlines potential directions for future research, particularly in improving model generalization and scalability. Furthermore, it highlights the broader societal and environmental impacts of precise soil nutrient prediction, particularly in supporting sustainable agriculture, enhancing food security, and promoting responsible resource management. By bridging technological innovation with practical applications, this chapter underscores the long-term value of spectral approaches for both scientific advancement and societal well-being.

Chapter 2

Technical Background of Remote Sensing for Soil Nutrient Prediction

This chapter explores the fundamentals of spectral data in remote sensing. It covers the electromagnetic spectrum, spectral reflectance principles, and the differences between multispectral and hyperspectral data. It introduces key sensor platforms and explains how these platforms acquire spectral images. The chapter also presents preprocessing techniques—such as noise removal, normalization, and feature selection—and highlights common challenges like noise, redundancy, and environmental variability. Understanding these fundamentals is crucial for accurately interpreting remote sensing data. Finally, it sets the stage for applying advanced modeling techniques to extract meaningful insights from spectral data.

2.1 Remote Sensing

Remote sensing is the science and technology of obtaining information about objects, areas, or phenomena without direct physical contact, by detecting and measuring Electromagnetic Radiation (EMR) reflected, emitted, or scattered from them [71]. The process begins with an energy source—natural (the Sun) or artificial (radar, LiDAR)—emitting electromagnetic energy. When this energy interacts with the Earth's surface, different materials absorb, reflect, and emit energy in distinct ways, producing unique patterns of electromagnetic response specific to each material [94]. Sensors

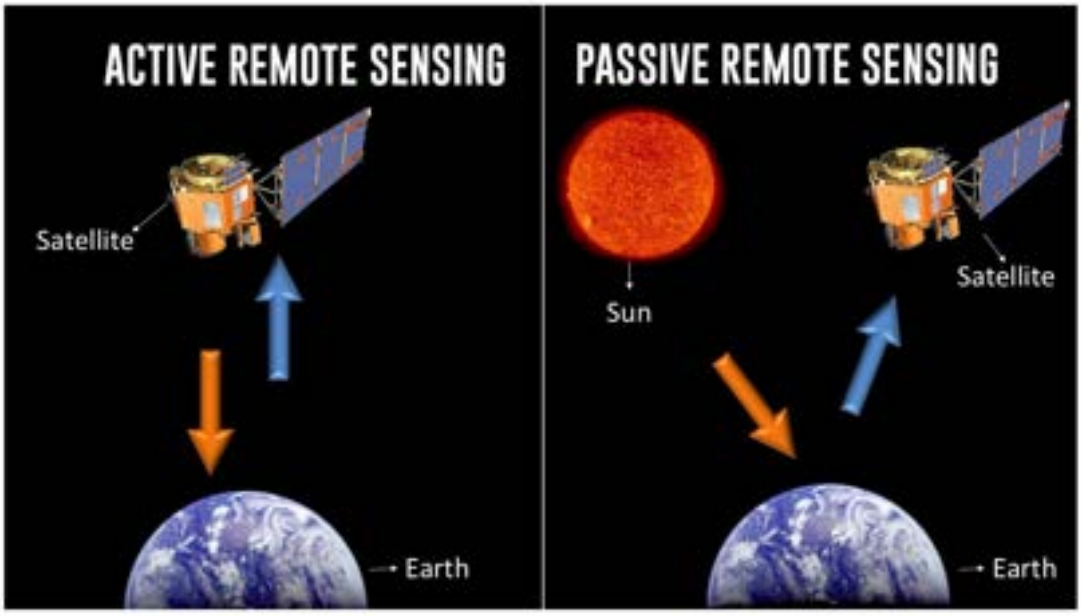

Figure 2.1: Active and passive remote sensing [11]

capture these signals, measure energy within defined wavelength ranges, transmit the data to receiving stations, and process it to extract meaningful information.

2.1.1 Types of Remote Sensing

Remote sensing broadly falls into two categories- active and passive, based on the source of electromagnetic energy used for observation. Figure 2.1 provides a conceptual overview of these two primary types of remote sensing. These methods are crucial in acquiring information about the Earth's surface and atmosphere without direct contact. Such methods form the basis of modern remote sensing applications in agriculture, environmental monitoring, and resource management.

1. ***Active Remote Sensing:*** It emits energy toward a target and measures the reflected or backscattered signal [95]. Examples include Synthetic Aperture Radar (SAR), Light Detection and Ranging (LiDAR), and radar altimeters. These systems operate independently of sunlight and can collect data day or night, even through cloud cover or vegetation.

2. ***Passive Remote Sensing:*** Instead of emitting a signal, it detects natural energy—primarily solar radiation—that is reflected or emitted by the Earth's surface [96]. Passive sensors are commonly employed in optical, thermal,

and infrared imaging applications. Passive optical remote sensing specifically captures the reflected portion of solar radiation across various wavelengths in the electromagnetic spectrum [97]. As a result, spectral data represent the surface reflectance across different electromagnetic spectrum bands. The amount of reflectance at each wavelength is influenced by the material's composition, structure, and moisture content, making it a valuable tool for surface characterization. Reflectance data is taken as input for various remote sensing applications, including vegetation monitoring, soil nutrient prediction, and land cover classification.

2.1.2 Applications of Remote Sensing

Remote sensing has a wide range of applications across various domains, enabling the collection and analysis of critical information without direct contact with the target area. Some of the major application areas include [98]:

1. **Agriculture** – Remote sensing plays an essential role in modern agriculture by allowing farmers and researchers to monitor crop conditions across large areas [99]. It helps identify different growth stages, detect plant stress caused by pests, drought, or nutrient deficiencies, and estimate crop yield [100]. These insights support precision farming by enabling efficient water use, timely fertilizer application, and effective pest control strategies, ultimately increasing productivity and sustainability [101].

2. **Forestry** – Remote sensing aids in tracking forest cover, detecting deforestation and illegal logging, and evaluating forest health [102, 103]. It also helps to classify forest types, estimate timber volume, and monitor regrowth after disturbances, such as wildfires [104]. Governments and conservation groups use satellite imagery for forest management, biodiversity protection, and carbon stock prediction to address climate change goals [105].

3. **Water Resources** – Remote sensing supports monitoring water bodies such as rivers, lakes, and reservoirs [106]. It helps detect seasonal changes in water levels, as well as drought-affected regions and flood-prone zones. These resources also

play a role in water quality monitoring by identifying polluted zones and tracking algal blooms [107]. The applications are critical for effective water resource planning, disaster preparedness, and ensuring water security [68].

4. **Soil and Land Management** – Satellite and aerial imagery provide valuable information for assessing soil conditions, identifying erosion-prone areas, and supporting land-use planning [108]. Remote sensing helps track land degradation, desertification, and improper agricultural practices [109]. It contributes to sustainable development by guiding policy decisions on soil conservation and land rehabilitation [110].

5. **Urban Planning** - Remote sensing enables detailed land use and land cover mapping in urban areas [111]. It helps monitor city expansion, infrastructure development, and environmental impacts of urbanization. Urban planners use this data to identify unauthorized constructions, manage transportation networks, plan new settlements, and conserve green space. Historical satellite records also help analyze urban growth trends over time [112, 113].

6. **Disaster Management** – Remote sensing is essential for early warning systems, real-time monitoring, and damage assessment during disasters like floods, earthquakes, cyclones, wildfires, and landslides [114, 115]. It helps authorities visualize the extent of damage, plan evacuations, and coordinate relief efforts. This also aids in post-disaster recovery by assessing land condition and rebuilding needs [116].

7. **Climate and Weather Studies** – Satellites monitor atmospheric patterns, cloud movement, rainfall distribution, snow and ice cover, and temperature variations [117, 118]. Long-term satellite records support climate change research by detecting trends in global warming, sea-level rise, and shifting weather patterns [119]. These insights are crucial for policymakers, researchers, and environmentalists.

8. **Environmental Monitoring** – Remote sensing is widely used for tracking changes in natural ecosystems. It supports the detection of pollution sources,

monitoring of wetlands, identification of habitat loss, and protection of endangered species [120, 121]. This also assists in assessing the environmental impact of industrialization, mining, and infrastructure development projects [122].

9. **Geology and Mining** – Geological surveys use remote sensing to map terrain features, identify rock formations, locate faults, and assess geohazards such as landslides or volcanic activity [123, 124]. In mining, it supports exploration of mineral resources and monitoring land use around mining sites, helping reduce environmental damage and ensure safe practices [125].

10. **Oceanography** – Remote sensing allows continuous observation of oceanic conditions such as currents, wave height, sea surface temperature, and coastal erosion [126,127]. It helps map coral reefs, monitor marine pollution, and support sustainable fisheries. These applications are critical for marine conservation and coastal zone management [128].

11. **Defence Sector** - In the defence sector, remote sensing is extensively used for surveillance, reconnaissance, and strategic planning [129]. It provides real-time images for border monitoring, target detection, and tracking troop movements [62]. High-resolution satellite imagery enhances situational awareness, supports intelligence gathering, and enables precise mission execution [130].

2.1.3 Electromagnetic Spectrum

The Electromagnetic Spectrum (EMS) refers to the entire range of electromagnetic radiation, classified according to wavelength or frequency [74]. The various components of the electromagnetic spectrum, in order of increasing wavelength and decreasing energy, include gamma rays, X-rays, ultraviolet (UV) rays, visible light (VIS), infrared (IR) (which includes Near-Infrared (NIR), Short-wave Infrared (SWIR), and Mid-Infrared (MIR)), microwaves, and radio waves [75]. Each region interacts differently with matter and serves unique applications in science and technology [76]. In soil analysis and nutrient prediction, only specific portions of the EMS—primarily the visible and infrared regions (especially NIR and SWIR)—are applicable [131]. These regions are sensitive to key nutrients, including organic matter, moisture content,

Table 2.1: Spectral regions, wavelength ranges, and common remote sensing applications.

Spectral Region	Wavelength Range	Common Applications
Ultraviolet (UV)	10 – 400 nm	Detection of mineral fluorescence; limited use due to atmospheric absorption
Visible (VIS)	400 – 700 nm	Human vision range; plant pigments (e.g., chlorophyll) absorb in blue and red
Near Infrared (NIR)	700 – 1300 nm	Sensitive to vegetation structure and soil moisture
Short-wave Infrared (SWIR)	1300 – 2500 nm	Rich in mineral and moisture information; key for soil nutrient prediction
Mid Infrared (MIR)	2500 – 25000 nm	Used in thermal imaging; captures emitted radiation rather than reflected light

and mineral composition. By analyzing how soil reflects or absorbs electromagnetic radiation in these bands, researchers can non-destructively estimate nutrient levels, supporting precision agriculture and soil health monitoring [57].

2.1.3.1 Relevant Spectral Regions in Remote Sensing:

Table 2.1 presents the key regions of the EMS that are widely used in environmental monitoring, along with the corresponding wavelength ranges and typical applications. Each spectral region provides unique information about the Earth's surface and atmosphere, making them critical for designing and interpreting remote sensing systems [74]. For example, the UV region can reveal mineral fluorescence, while the VIS range corresponds to human vision and is essential for studying vegetation pigments [76]. The NIR range is susceptible to vegetation structure and soil moisture, whereas the SWIR captures valuable mineralogical and moisture-related information, supporting applications such as soil nutrient assessment. The MIR range, on the other hand, is used for thermal imaging to measure emitted radiation, enabling surface temperature mapping. By integrating these regions, spectral data systems can record hundreds of narrow bands, allowing precise material identification and supporting a broad range of environmental monitoring tasks [57].

2.1.3.2 What are Spectral Bands?

Spectral bands refer to specific, contiguous wavelength intervals within the electromagnetic spectrum where remote sensing sensors record reflectance or radiance data [69].

Each band is designed to capture energy within a defined range of wavelengths, typically expressed in nanometers (nm) or micrometers (μm), corresponding to different spectrum regions such as VIS, NIR, SWIR, or TIR. The reflectance values measured in these bands provide insights into the physical and chemical characteristics of the Earth's surface, as materials interact differently with electromagnetic radiation depending on composition and structure [5]. By analyzing the variation in reflectance across multiple bands, it is possible to identify and distinguish surface materials, monitor vegetation health, assess water quality, and detect soil nutrients. The selection and number of spectral bands influence the sensor's ability to capture fine spectral details, making them a critical factor in the accuracy of spectral analysis and the scope of remote sensing applications [132].

2.1.3.3 What is Spectral Reflectance?

Spectral reflectance refers to the proportion of incident electromagnetic radiation reflected by a surface as a function of wavelength [77]. It is expressed as a unitless ratio, typically ranging from 0 (no reflection) to 1 (total reflection), and varies according to the physical and chemical properties of the material, surface roughness, and moisture content [90]. This wavelength-dependent variation in reflectance produces a spectral signature unique to each material, enabling its identification in remote sensing applications.

In practice, remote sensing instruments do not directly measure reflectance. Instead, these instruments capture radiance—the energy received by the sensor—often recorded as raw Digital Number (DN) values. These measurements are influenced by atmospheric scattering, absorption, and sensor calibration. The raw data undergo radiometric correction (adjusting for sensor and illumination effects) and atmospheric correction (removing atmospheric interference) to retrieve accurate reflectance values. The resulting spectral reflectance data form the foundation for various remote sensing analyses, including vegetation health monitoring, soil nutrient prediction, and land-cover classification [133].

2.1.3.4 What are the Spectral Signatures of Earth Surface Materials?

A spectral signature is the unique pattern of reflectance exhibited by a material over a range of wavelengths [78]. Each material reflects, absorbs, and transmits electromagnetic radiation differently, resulting in distinctive spectral curves when reflectance is plotted against wavelength. These spectral signatures are essential for identifying and distinguishing between surface features such as vegetation types, soil compositions, and water bodies.

For example, healthy vegetation strongly absorbs radiation in the blue (450 nm) and red (670 nm) regions due to chlorophyll. However, it reflects the NIR region (700–1300 nm) significantly due to the internal leaf structure. Because of high absorption, water bodies typically show low reflectance in the visible and NIR regions. Dry soils generally have higher reflectance in the SWIR region than moist soils. By analyzing these patterns, remote sensing systems can classify and monitor Earth surface features based on spectral behavior [134].

In soil spectroscopy, long-range spectral dependencies can arise when widely separated wavelength regions within a spectral signature jointly contribute to predicting a soil property [135]. For example, soil organic carbon can influence the spectral signature across multiple wavelength regions, with information from one region complementing that from another. Such relationships motivate the use of models capable of capturing long-range dependencies in spectral signatures [135, 136].

2.1.4 What is Spectral Data?

Spectral data has become indispensable in modern remote sensing and environmental monitoring. It provides a non-invasive, scalable, and efficient means of analyzing the Earth's surface by measuring how different materials interact with electromagnetic radiation [80]. These interactions—primarily in the form of reflectance, absorbance, or transmittance—allow us to detect and distinguish between various natural and artificial features, such as soil types, vegetation cover, water bodies, and urban infrastructure [137]. When radiation interacts with a surface, it can undergo three fundamental processes: reflection, absorption, and transmission. How a material responds to radiation across different wavelengths defines its spectral signature, which

is critical for identifying and analyzing surface materials in remote sensing [138].

- ***Reflectance (R(λ)):*** It is the fraction of incident radiation reflected by the surface at wavelength λ.
- ***Transmittance (T(λ)):*** It is the fraction of radiation transmitted through the surface at wavelength λ.
- ***Absorbance (A(λ)):*** It is the fraction of radiation absorbed by the material and converted to other forms of energy, usually heat, through the surface at wavelength λ.

The energy conservation equation 2.1 relates these three processes:

$$R(\lambda) + A(\lambda) + T(\lambda) = 1 \tag{2.1}$$

In most remote sensing applications, transmittance is often negligible, especially for opaque surfaces. It simplifies the relationship as shown in equation 2.2:

$$R(\lambda) + A(\lambda) \approx 1 \tag{2.2}$$

This simplified form is commonly used for modeling and interpreting surface reflectance and absorption in passive optical remote sensing, where sensors rely on solar illumination to detect reflected radiation from the Earth's surface.

In remote sensing, the quality and usefulness of the data collected depend on various kinds of resolution—namely, spatial, spectral, radiometric, and temporal resolution [139]. However, the actual values are sensor-dependent and often influenced by the type of electromagnetic radiation used.

1. ***Spatial Resolution:*** It refers to the size of the smallest object that can be detected on the ground, often described by the size of a pixel (e.g., 10 meters or 30 meters). The clarity of features on Earth's surface depends on the size of the pixel and the number of pixels in a given image.
2. ***Temporal Resolution:*** It indicates how frequently a sensor can revisit the same area over time. Temporal resolution is crucial for understanding the direction and extent of change in the study area.

3. ***Radiometric Resolution:*** It refers to the ability of a sensor to capture the minute differences in the radiated energy from the Earth's surface. Radiometric calibration relates to the sensor's ability to detect subtle differences in energy or brightness measured in bits (e.g., 8-bit, 12-bit).

4. ***Spectral Resolution:*** It describes a sensor's ability to distinguish between different electromagnetic spectrum wavelengths, which is most relevant to this work. If a sensor can detect more wavelengths of the electromagnetic spectrum, it will capture greater details of land use and cover.

This thesis focuses on spectral resolution, which enables the capture of narrowly spaced wavelength intervals and is crucial for identifying subtle spectral variations that are significant for environmental monitoring and analysis. Spectral data may be multispectral, with tens of broad wavelength bands, or hyperspectral, with hundreds of narrow and contiguous bands. A sensor with high spectral resolution captures data in very narrow and closely spaced wavelength intervals, enabling detection of small but significant differences in material properties [98]. Each material exhibits a unique spectral fingerprint across these fine bands in hyperspectral data, allowing accurate identification and classification [140]. A more detailed discussion of these spectral data types is provided in section 2.2. In contrast, sensors with low spectral resolution—such as those used in basic RGB (Red, Green, Blue) data—capture fewer and broader bands, which may be insufficient for detailed material analysis. These data, represented as reflectance values across spectral bands, form the basis for analyzing material properties. It also supports applications such as vegetation health assessment, soil nutrient prediction, and land cover classification. Understanding spectral resolution is therefore crucial, as it directly impacts the accuracy and depth of spectral data interpretation for both scientific and practical applications.

2.1.4.1 Significance of Spectral Reflectance in Remote Sensing

A common source of confusion in spectral data analysis is the misunderstanding regarding what exactly is captured by remote sensing sensors: reflectance, absorbance, or transmittance? In passive optical remote sensing, the data acquired by the sensors represents only the reflected portion of incoming solar radiation — that is, spectral

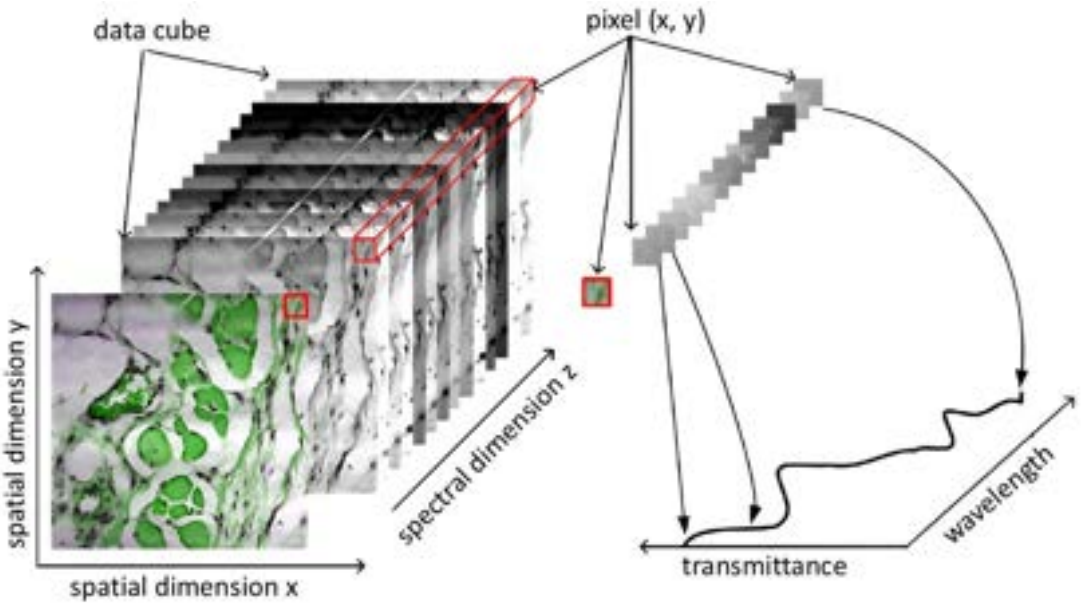

Figure 2.2: Spectral imaging showing three-dimensional spectral cube with spatial dimensions (x, y) and spectral dimension (z), each pixel holding a full spectral profile [12].

reflectance [96]. The absorbed and transmitted components are not directly measured.

While all three processes—reflection, absorption, and transmission—occur when light interacts with a surface, only the reflectance is detected and recorded by the sensor [69]. Absorbance and transmittance are either estimated through models or inferred based on known material characteristics. Therefore, when interpreting spectral data (such as hyperspectral curves or band-wise reflectance), it is crucial to understand that these values correspond exclusively to the reflectance behavior of the surface.

2.1.5 What is Spectral Imaging?

Spectral data can be captured in multiple forms, such as direct reflectance measurements from a single point or area, or as images through spectral imaging. Spectral imaging records image data across multiple, narrow wavelength bands of the electromagnetic spectrum, allowing detailed characterization of surface materials [12]. Passive optical remote sensing captures reflected solar radiation at different wavelengths, producing spectral data for each pixel in the image.

Figure 2.2 illustrates the spectral imaging that is typically organized into a three-dimensional spectral cube, where the two spatial dimensions (x, y) represent the location

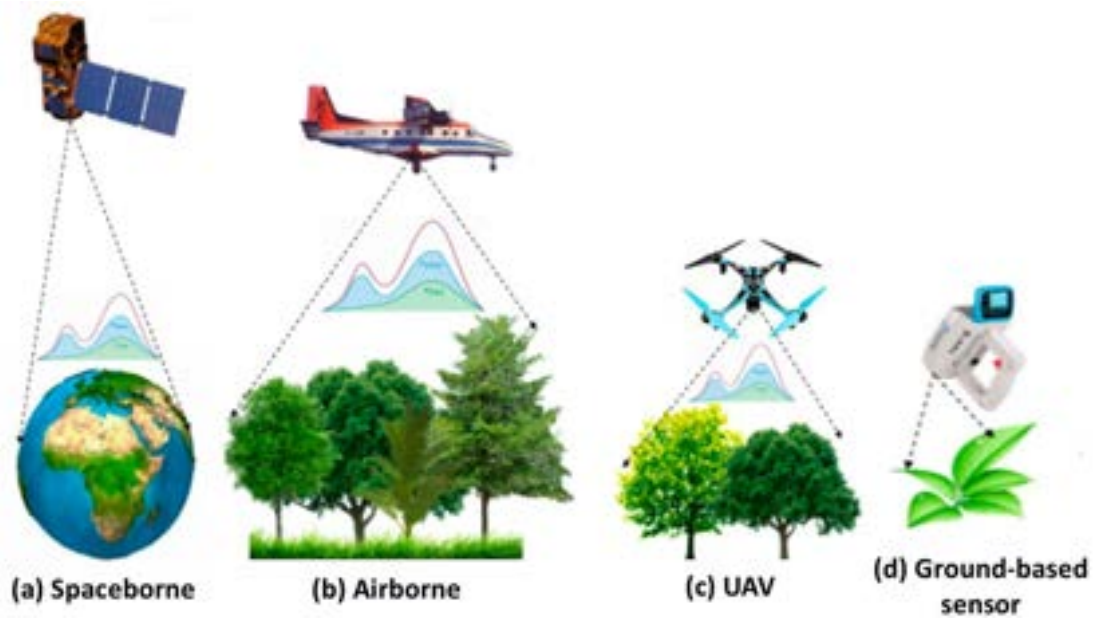

Figure 2.3: Sensors used to capture spectral data [13]

of pixels on the ground, and the third dimension (z) contains the spectral reflectance or radiance values at each wavelength. Unlike conventional RGB data, which is limited to three broad bands (red, green, and blue), spectral imaging can encompass dozens (multispectral) to hundreds (hyperspectral) of narrow, contiguous bands, enabling highly detailed discrimination of materials with similar visual appearances. Each material has a unique spectral signature—a distinct pattern of reflectance or absorption across wavelengths—making it possible to identify and quantify specific substances [61]. This capability is crucial for various applications, including vegetation health monitoring, soil nutrient prediction, mineral mapping, water quality assessment, and detecting environmental changes.

2.1.6 Acquisition of Spectral Data and Images: Sensors and Platforms

The acquisition of spectral data relies on various sensor platforms, each suited for specific spatial scales, resolutions, and application contexts. Figure 2.3 shows the different types of sensors and platforms used to capture the spectral data. These platforms are generally classified as satellite, airborne, UAV-based, and ground-based, each offering unique advantages and limitations [141]. Ground-based instruments,

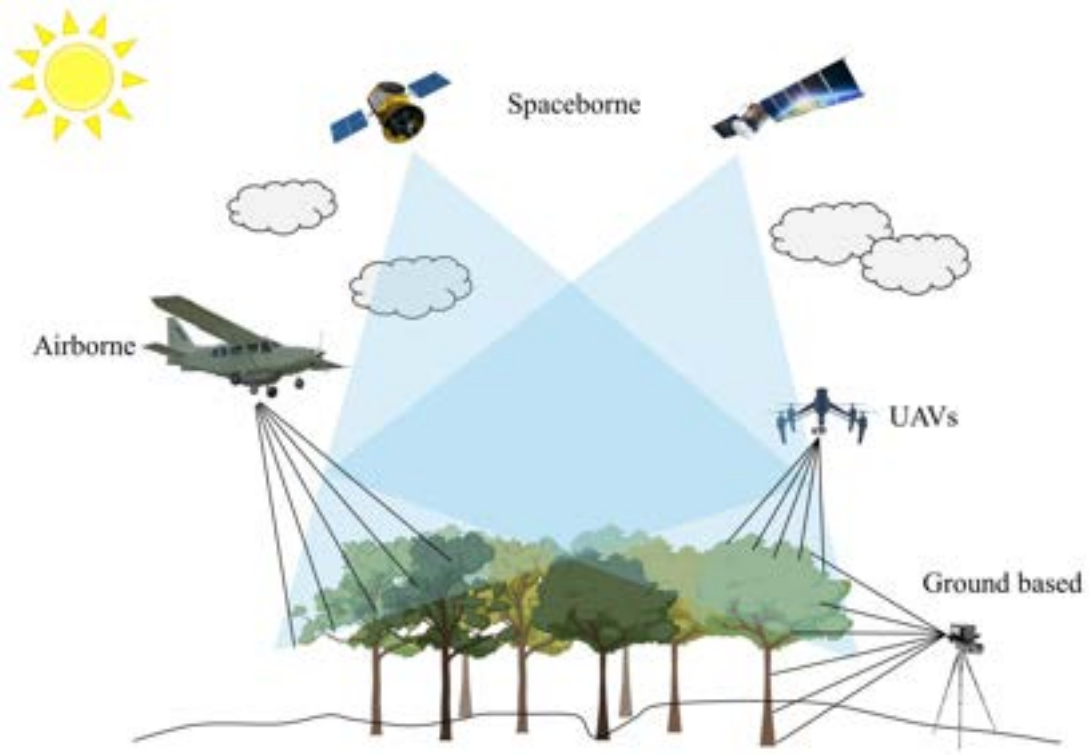

Figure 2.4: Soil spectral data acquisition using remote sensing platforms [14]

such as spectroradiometers, capture direct spectral data (point measurements without imagery), while satellites, airborne platforms, and UAVs capture spectral data in spectral images. Figure 2.4 summarizes how different sensing platforms contribute to soil spectral data acquisition for AI-based nutrient prediction models. Understanding the capabilities is essential for selecting appropriate data sources for remote sensing tasks. These sensors and platforms are broadly categorized into:

1. ***Spaceborne Sensors:*** The remotely sensed data from these sensors have immense potential as an enabling tool for generating spatial maps of the upper horizon, owing to the proven background in interlinkages among electromagnetic radiation, such as Sentinel-2, Landsat series, Hyperion (EO-1), EnMAP, and PRISMA [7]. Table 2.2 outlines the key strengths and limitations of spaceborne sensors. These sensors provide extensive spatial coverage and frequent revisit times, making them highly suitable for global monitoring and time-series analysis. The data is radiometrically calibrated and widely used for quantitative applications. However, the high orbital altitude limits the lower spatial resolution of spaceborne platforms, and data acquisition can be affected by atmospheric

Table 2.2: Advantages and limitations of spaceborne sensors [7]

Advantages	**Limitations**
1. Provides large-area coverage and frequent revisits.	1. Suffers from lower spatial resolution due to high orbit altitude.
2. Offers long-term, continuous global datasets.	2. Often obstructed by clouds and atmospheric effects.
3. Radiometrically calibrated sensors enable quantitative analysis.	3. Delayed data acquisition and lower flexibility for tasking.

conditions, such as cloud cover. Additionally, tasking flexibility is restricted, as most satellites follow predefined acquisition schedules.

Characteristics: Spaceborne sensors acquire data over vast geographical areas at regular intervals, typically with revisit times ranging from 5 to 16 days. The spectral range varies significantly; for example, Sentinel-2 captures 13 bands spanning the visible to SWIR regions, while Hyperion records 220 hyperspectral bands. These sensors provide publicly available and commercial datasets, with data often preprocessed by space agencies to include atmospheric correction and geometric alignment, ensuring accuracy and usability.

2. ***Airborne Sensors:*** These sensors offer the ability for spatial assessment, providing a more accurate mapping of the variability observed. The information can cover large areas, even for a single flight mission, since aircraft provide adequate flight duration. For example, human-crewed aircraft with imaging spectrometers, such as AVIRIS, CASI, and HyMap [7]. Table 2.3 presents the advantages and limitations of aircraft-based platforms. These systems offer high spatial resolution and customizable sensor configurations, making them suitable for scientific campaigns and regional surveys. The flexibility allows rapid deployment and on-demand data collection. However, aircraft-based operations are expensive, logistically complex, and limited in coverage area per flight. Regulatory approvals and sensor integration also require careful planning, making them less practical for routine, large-scale monitoring.

   Characteristics: Airborne sensors operate at altitudes between UAVs and satellites, typically ranging from a few hundred meters to several kilometers. These sensors can carry heavy and sophisticated hyperspectral sensors that are not

Table 2.3: Advantages and limitations of airborne sensors [7]

Advantages	**Limitations**
1. Capable of very high spatial resolution.	1. Typically covers a limited area per flight, making wide coverage costly.
2. Flexible deployment with customizable sensor setup.	2. Complex logistics and regulatory barriers (e.g., flight permissions).
3. Rapid sensor reconfiguration and maintenance are possible.	3. Operational and per-area costs are high.

feasible to mount on UAVs, enabling wide-area, high-resolution data collection with greater flexibility than satellites. These platforms are often deployed in campaigns coordinated by research institutions or government agencies for targeted environmental and resource monitoring.

3. ***Unmanned Aerial Vehicles (UAVs):*** It combines characteristics of spaceborne and airborne platforms, offering a unique capability to provide the resolution needed to cover the diversity of agricultural landscapes. In addition, the ever-increasing analytical capabilities for data handling could potentially lead to less time-consuming image interpretation [8]. UAV-mounted hyperspectral or multispectral sensors such as MicaSense RedEdge and Headwall Nano-Hyperspec. Table 2.4 summarizes the advantages and limitations of UAVs in remote sensing. UAVs provide ultra-high-resolution imagery with flexible deployment over small areas, making them ideal for precision agriculture, soil surveys, and close-range environmental monitoring. These drones are more affordable than manned aircraft and are less affected by cloud cover. However, usage is constrained by battery limitations, regulatory restrictions such as line-of-sight operations, and a limited spatial footprint. Furthermore, UAV-based data often requires extensive post-processing and storage resources.

   Characteristics: UAVs offer a flexible, scalable, and user-friendly platform for capturing spectral data. Operating at low altitudes, typically between 20 and 150 meters, these drones provide high spatial and temporal resolution, making them suitable for detailed monitoring. UAVs are equipped with Global Positioning System (GPS) and Inertial Measurement Unit (IMU) sensors to enable precise georeferencing, ensuring accurate spatial alignment of the collected

Table 2.4: Advantages and limitations of UAVs [8]

Advantages	**Limitations**
1. Provide ultra-high spatial and temporal resolution data.	1. Limited operational range due to battery life and regulatory restrictions.
2. Highly flexible and affordable compared to manned aircraft.	2. Operational scalability is constrained for large area surveys.
3. Less atmospheric interference; data acquisition over complex terrain.	3. Data processing pipelines can be complex and resource-intensive.

data. Depending on the payload capacity, UAV's can carry multispectral and hyperspectral sensors, making them versatile for various remote sensing applications.

4. ***Ground-Based Sensors:*** These instruments measure atmospheric and environmental conditions in situ, such as temperature, humidity, and spectral reflectance. The spectroradiometers, handheld hyperspectral sensors, and field spectrometers are various forms of the ground-based sensors [7]. Table 2.5 presents the advantages and limitations of ground-based sensors in remote sensing. These systems offer precise, site-specific data and are critical for validating satellite and airborne observations. These sensors are ideal for continuous monitoring at fixed locations, especially for field experiments and calibration studies. However, the spatial coverage is minimal, and widespread deployment is labor- and resource-intensive. Ground-based sensors are not feasible for broad-scale assessments but remain essential for ensuring the accuracy of remotely sensed data.

   Characteristics: Ground-based sensors offer high spectral and radiometric accuracy, enabling the capture of detailed reflectance data for precise analysis. These sensors are commonly employed in field campaigns and laboratory settings, with instruments generally portable or mounted on tripods, making them suitable for accurate point-based measurements. Although the spatial coverage is limited, ground-based sensors are ideal for micro-scale studies and serve as valuable tools for validating data from airborne and spaceborne platforms.

Table 2.5: Advantages and limitations of ground-based sensors [7]

Advantages	**Limitations**
1. Provide highly accurate, localized validation data (ground truth).	1. Only cover minimal spatial areas.
2. Useful for calibration and validation of airborne or satellite data.	2. Resource- and labor-intensive to deploy and maintain.
3. Ideal for high-detail measurements and temporal consistency at small sites.	3. Not suitable for broad-area or long-term continuous monitoring.

2.1.6.1 Calibration and Correction

Accurate spectral data acquisition relies on proper calibration and correction techniques to eliminate the effects of instrument response, atmospheric interference, and background noise. The following steps are essential to ensure the integrity and comparability of spectral measurements across different sensors, platforms, and periods.

1. ***Radiometric Calibration:*** It converts raw sensor digital numbers into physical units (e.g., radiance or reflectance) [142]. This calibration accounts for sensor-specific responses, allowing for accurate comparisons across sensors and datasets. It uses pre-flight sensor characterization or in-flight onboard calibration sources. It ensures that sensor readings correspond to absolute or relative light intensities.

2. ***Atmospheric Correction:*** It removes the effects of atmospheric gases, aerosols, and water vapor on the recorded spectra [142]. Standard techniques include the empirical line method, which uses ground reflectance targets to match recorded and known values. Radiative Transfer Models (e.g., MODTRAN, 6S) simulate atmospheric effects and invert to surface reflectance. Dark Object Subtraction (DOS) assumes dark pixels should reflect minimal light and adjusts accordingly. It is essential for satellite and airborne data, where atmospheric conditions can significantly alter spectral signatures.

3. ***Dark Current and White Reference Correction:*** It addresses noise caused by electronic activity in the sensor when no light is entering. Measured by closing the shutter, using covered pixels, and subtracting the background signal. White reference correction uses a calibrated white reference panel (e.g., Spectralon) that

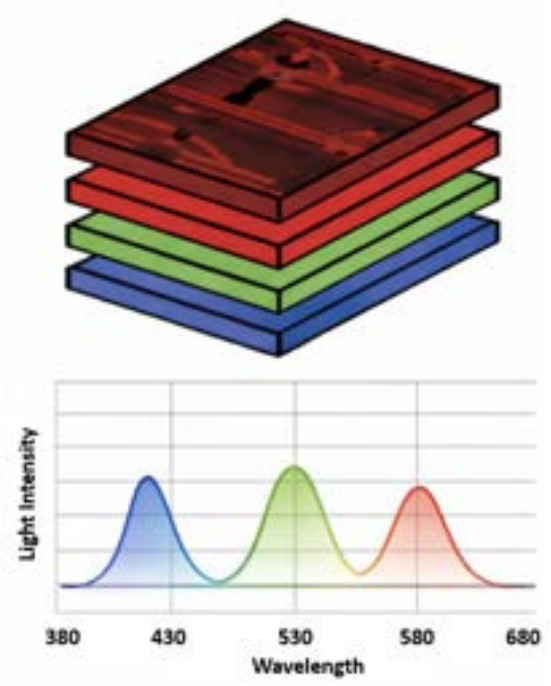

Figure 2.5: Multispectral data bands and the spectral response [15]

reflects nearly all incoming light. Normalization of a target's reflectance spectra involves dividing the target reflectance by the reference reflectance. It ensures that the reflectance values are relative and not influenced by varying illumination conditions [143].

These calibration and correction procedures are foundational for producing accurate, consistent, and comparable spectral datasets, which are critical for reliable remote sensing analysis and interpretation.

2.2 Types of Spectral Data

The major types of spectral data currently used for various applications are multispectral and hyperspectral. The scientific community is also working towards developing ultraspectral and super-ultraspectral resolutions, which will help even more than hyperspectral contiguous wavelength bands. However, these data types are currently in the development stages, and no data is available for field applications [144].

2.2.1 Multispectral Data

Multispectral Data (MSD) captures the data at a few selected and discrete spectral bands across the electromagnetic spectrum [81]. This technique provides valuable information about the Earth's surface by measuring reflected light in several broad wavelength regions [145]. MSD sensors are commonly found on satellites, drones, and handheld devices, making them highly accessible and widely used in environmental, agricultural, and urban applications. Figure 2.5 illustrates the working principle of a multispectral data system, which captures surface reflectance data across a limited number of discrete spectral bands. Each layer in the top schematic represents a sensor or filter dedicated to a specific wavelength range, such as blue (around 430–480 nm), green (around 530 nm), and red (around 630–680 nm)

2.2.1.1 Key Characteristics

1. ***Number of Bands:*** Multispectral sensors typically record data in 3 to 13 discrete bands. These bands are selected based on the sensitivity to key features in the visible and near-infrared portions of the spectrum. Standard band combinations include RGB and NIR. These bands are generally broad and separated by wavelength intervals.

2. ***Spectral Resolution:*** MSD provides moderate spectral resolution, meaning each band covers a broader range of wavelengths. It limits the ability to detect subtle spectral variations, but it is sufficient for distinguishing between broad categories of materials such as vegetation, water, and soil.

3. ***Spatial Resolution:*** MSD often offers high spatial resolution compared to hyperspectral systems. It enables the identification of fine-scale features in landscapes, making MSD suitable for detailed mapping and analysis.

4. ***Spectral Band Characteristics:*** The bands used in MSD are non-contiguous, meaning these bands do not form a continuous spectrum. Instead, it targets specific wavelengths that provide useful information, such as chlorophyll absorption in the red band and vegetation reflectance in the NIR.

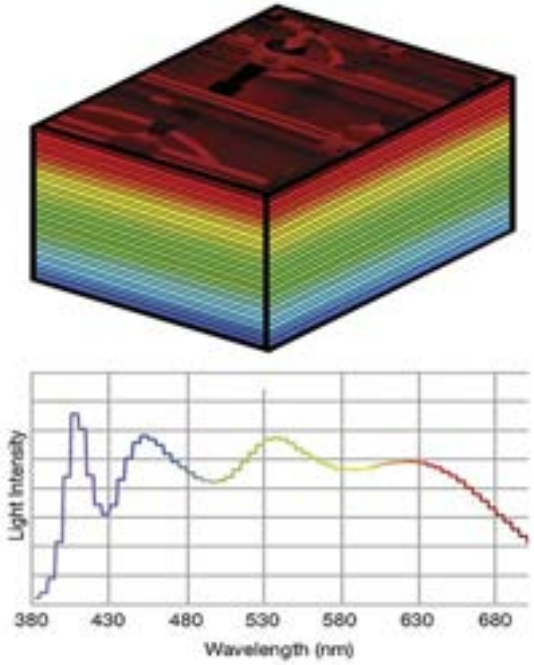

Figure 2.6: Hyperspectral data bands and the spectral response [15]

5. ***Common Platforms and Sensors:*** Widely used sensors include the Landsat series (e.g., Landsat-8 with 7 or 11 bands), Sentinel-2, and MODIS, among others. These platforms provide long-term and high-resolution data that support a range of Earth observation activities.

2.2.2 Hyperspectral Data

Hyperspectral Data (HSD) is an advanced data type that captures the full reflectance spectrum [82]. Unlike MSD, which samples a few broad spectral bands, HSD measures hundreds of narrow, contiguous spectral bands, providing a continuous spectral profile that can uniquely characterize different materials based on the spectral signatures [146]. Figure 2.6 illustrates the structure and output of a hyperspectral data system. The top part of the figure shows a sensor capturing data across dozens to hundreds of narrow, contiguous spectral bands, covering a wide range of wavelengths. The bottom graph displays a continuous spectral reflectance curve, representing the intensity of reflected light across wavelengths from approximately 380 nm to 680 nm.

2.2.2.1 Key Characteristics

1. ***Number of Bands:*** HSD sensors typically capture hundreds to thousands of spectral bands, allowing the detection of narrow absorption features associated with specific chemical or mineral properties. This high dimensionality facilitates detailed material discrimination and enhances the ability to model subtle spectral variations.

2. ***Spectral Resolution:*** HSD offers high spectral resolution, often with bandwidths of 5–10 nm or less. This high granularity enables the discrimination of spectrally similar materials. Such a resolution is particularly beneficial for identifying soil constituents, assessing crop health, and determining mineral content.

3. ***Spectral Continuity:*** Unlike MSD, HSD bands are contiguous, spanning a continuous range of wavelengths without gaps. This continuity is crucial for detailed spectral analysis. It supports the extraction of complete spectral signatures and improves the accuracy of feature-based classification and regression.

4. ***Spatial Resolution:*** HSD generally provides lower spatial resolution than MSD due to the sensor design trade-offs. However, advances in sensor technology are helping to alleviate this limitation. Emerging sensors aim to strike a balance between spatial and spectral resolution to meet the diverse needs of remote sensing applications.

5. ***Common Platforms and Sensors:*** Prominent HSD sensors include AVIRIS (Airborne Visible/Infrared data Spectrometer), Hyperion, HyMap, and PRISMA. These sensors are used in airborne and satellite missions. Each platform offers varying capabilities in terms of spatial coverage, revisit frequency, and application domain.

2.2.3 Difference between Hyperspectral and Multispectral Data

Hyperspectral and multispectral data methods rely on taking and examining data at various wavelengths; these two strategies have important distinctions. Engineers and

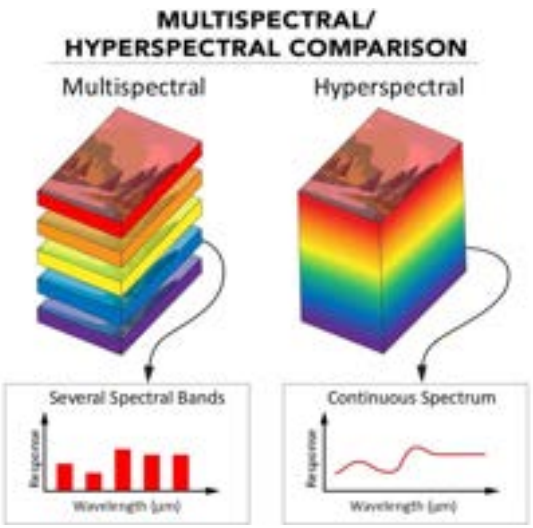

Figure 2.7: Comparison of the image stacks in multispectral and hyperspectral data. [16]

system designers who want to utilize spectrum data for applications must be aware of these distinctions. Figure 2.7 compares the image stacks in multispectral and hyperspectral data. Multispectral bands are discrete and non-overlapping segments of the electromagnetic spectrum. Contiguous spectrum refers to the continuous, fine-resolution capture of data across the entire electromagnetic spectrum, a defining characteristic of hyperspectral data.

Hyperspectral data allows for precise material identification due to the richness of spectral data, making it suitable for applications such as mineral exploration, crop stress detection, and soil nutrient prediction [147]. Multispectral data, in contrast, is more suited for broader, less detailed tasks, such as vegetation monitoring or land use classification, due to its limited number of bands [81]. Furthermore, hyperspectral data often require dimensionality reduction and complex algorithms, whereas multispectral data can be easily processed with conventional remote sensing techniques.

Table 2.6 compares the multispectral and hyperspectral data in the spectral, computational, and application-related characteristics. Multispectral sensors typically acquire data in 3–10 broad bands, offering lower spectral resolution and faster processing with simpler computational requirements. However, its limited spectral detail often leads to challenges in accurately identifying materials, especially when spectral signatures are similar, resulting in a higher risk of misclassification. In contrast, hyperspectral sensors capture over 100 narrow, contiguous bands, providing high spectral resolution that enables precise material differentiation, detection of subtle

Table 2.6: Difference between multispectral and hyperspectral data [9, 10]

Key Factor	Multispectral Data	Hyperspectral Data
Spectral Characteristics		
Number of Bands	3–10 distinct bands with broader wavelength coverage.	100+ contiguous narrow bands, capturing detailed spectral information (up to thousands).
Spectral Resolution	Limited spectral resolution, suitable for general analysis.	High spectral resolution, allowing better characterization of the chemical and physical properties.
Material Identification	Lower accuracy in differentiating materials with similar spectral signatures.	Enables accurate identification and classification of spectrally similar materials due to finer resolution.
Detection of Subtle Differences	Cannot detect minor spectral variations, leading to confusion in similar classes.	Detects subtle biochemical and environmental variations, enhancing class separability.
Processing and Computation		
Data Processing Speed	Faster processing due to smaller data size and fewer features.	Slower processing due to high dimensionality and large volume of data.
Processing Complexity	Simpler to process and analyze using standard algorithms.	Requires advanced algorithms and techniques due to the complexity of the data.
Preprocessing and Training Time	Faster due to smaller datasets and simpler structures.	Requires longer preprocessing and training due to noise removal and handling of dimensionality.
Storage and Computation Needs	Low storage and computational demands.	Requires significant storage capacity and computational power due to data size and complexity.
Risk of Overfitting	Lower risk of overfitting due to fewer input features.	Higher risk due to high dimensionality, requiring proper feature selection or regularization.
Application and Model Performance		
Model Performance	Limited performance in complex or high-precision tasks due to fewer features.	Superior performance in complex tasks by leveraging rich spectral detail.
Flexibility in Applications	Less flexible for precision or high-resolution tasks.	Highly flexible and supports advanced applications like chemical analysis, vegetation stress detection, etc.
Misclassification Risk	High risk of misclassification, meaning incorrect labeling of similar materials due to limited spectral detail.	Low risk when properly processed, as detailed spectral information helps distinguish spectrally close classes.
Data Availability and Cost		
Cost	Relatively cheaper and more accessible for general-purpose use.	Expensive, requiring specialized sensors and processing tools.
Data Availability	Widely available from multiple satellite and drone platforms.	Limited availability; mostly used in research and specialized applications.

biochemical variations, and improved model performance. Despite its advantages, hyperspectral data have increased cost, limited availability, complex preprocessing, and higher storage and computational needs. This comparison highlights the trade-off between the ease of access and processing in multispectral systems versus the analytical power and accuracy of hyperspectral systems.

2.3 Advantages of Spectral data

Spectral data are essential in remote sensing because these data allow for observing, analysing, and understanding surface materials and environmental conditions without physical contact [137]. The use of spectral data has become increasingly important due to its wide range of benefits, as outlined below:

1. ***Non-Destructive and Efficient Analysis:*** Spectral data enables the analysis of materials such as soil, vegetation, or water without physically disturbing them. It is particularly valuable in environmental science, where frequent or large-scale sampling can be invasive, expensive, and time-consuming. Spectral methods preserve the site's integrity while providing reliable information [148].

2. ***Rapid and Scalable for Large Areas:*** Unlike traditional field-based methods, spectral sensing—especially from satellites and drones—allows for rapid data collection across vast geographical regions. This scalability makes it practical for national or global monitoring programs and supports frequent updates for dynamic environments [149].

3. ***Identification Using Unique Spectral Signatures:*** Every material has a distinct way of reflecting or absorbing electromagnetic radiation across different wavelengths. These patterns, called spectral signatures, allow precise identification of materials—such as differentiating between soil types, crop species, or pollutants—based on the spectral response curves [66].

4. ***Enables Remote Sensing From Various Platforms:*** Spectral data can be collected from different platforms, including handheld instruments, Unmanned Aerial Vehicles (UAVs), aircraft, and satellites. This flexibility enables

researchers to collect data in remote, hazardous, or otherwise inaccessible areas, such as flood zones or mountainous regions [129].

5. ***Compatible With Automated, AI-Based Analysis:*** The high-dimensional nature of spectral data makes it suitable for machine learning and deep learning applications. Training models can automate environmental monitoring systems, enabling faster, more accurate analysis of soil and vegetation properties [150].

2.4 Challenges in Spectral Data Analysis

Despite its great potential, analysing spectral data presents several challenges that researchers and practitioners must address to ensure accurate and meaningful results [151]. The challenges in spectral data analysis are as follows:

1. ***High Dimensionality (Curse of Dimensionality):*** Hyperspectral data often contains hundreds of spectral bands, many of which may be redundant or irrelevant. This high dimensionality can lead to overfitting in machine learning models, increased computational costs, and difficulty in visualization [152]. Dimensionality reduction and feature selection methods (e.g., PCA, band selection algorithms) are often employed to mitigate this issue.

2. ***Noise and Redundancy in Spectral Bands:*** Spectral sensors may record noise due to atmospheric effects, sensor sensitivity, or environmental interference. Adjacent bands often carry similar information, leading to redundancy [83]. Preprocessing techniques, such as smoothing, filtering, and denoising, are necessary to enhance data quality.

3. ***Variability Due to Environmental Conditions:*** Spectral reflectance is influenced by varying illumination angles, moisture content, surface roughness, and time of day [83]. Environmental variability introduces inconsistency in spectral data collection across space and time. Normalization techniques and field calibration can help in minimizing these effects.

4. ***Sensor Calibration Issues:*** Inconsistent sensor calibration across platforms or over time can affect spectral fidelity. Calibration drift, sensor aging, and inter-

sensor variability must be accounted for [152]. Routine calibration using known references and inter-comparison with other datasets is essential.

Addressing these challenges is crucial for the reliable interpretation, modeling, and application of spectral data in real-world scenarios such as precision agriculture, mineral exploration, and ecosystem monitoring.

2.5 Automated Spectral Data Preprocessing Pipeline for Machine Learning Based Prediction

Preprocessing is vital in enhancing the quality and usability of spectral data before applying any machine learning models [153]. It ensures that the data is clean, consistent, and informative, essential for accurate and reliable predictions. Figure 2.8 illustrates the automated workflow for preparing spectral data before machine learning modeling. The process begins with converting raw spectral images into structured spectral data, followed by noise removal to eliminate unwanted distortions or sensor-related artifacts. This step is crucial to prevent the model from learning irrelevant patterns. Next, spectral normalization techniques are applied to bring all reflectance values to a comparable scale, thus improving data consistency across samples. Feature selection is then performed to identify and retain only the most informative spectral bands or features, reducing dimensionality and focusing the model on the most relevant information. Finally, the preprocessed data is passed into machine learning or deep learning models for prediction.

Each of these steps significantly improves model accuracy, reduces overfitting, and enhances interpretability. Without proper preprocessing, spectral data may contain noise, redundancy, and irrelevant features, leading to poor model performance and unreliable predictions [154]. Moreover, a well-designed preprocessing pipeline improves model efficiency and reduces computational time and resource requirements. Therefore, an automated and systematic preprocessing pipeline is essential for building robust, accurate, and interpretable models for spectral data prediction.

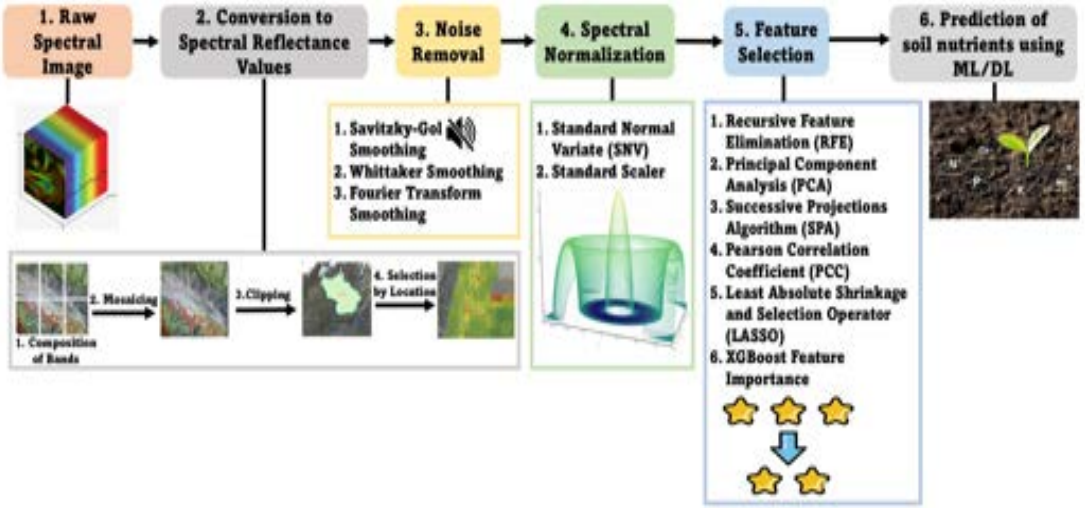

Figure 2.8: Automated workflow of preparing spectral data for ML models.

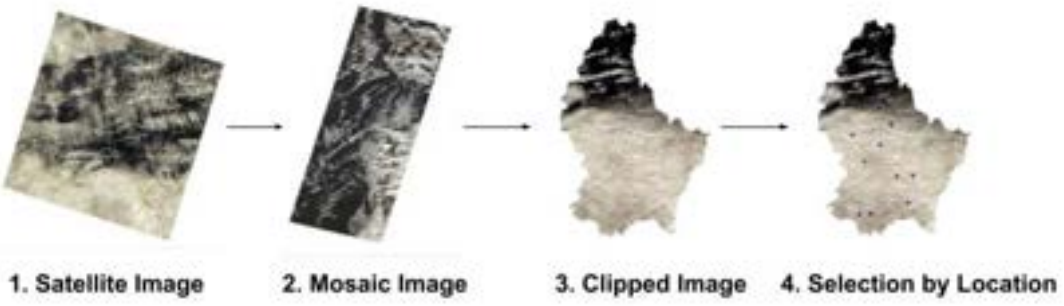

Figure 2.9: Satellite image preprocessing steps

2.5.1 Conversion of Raw Spectral Images into Spectral Reflectance Values

Transforming raw sensor outputs into structured, analyzable datasets suitable for ML models is crucial in spectral data preprocessing. Regardless of the sensor platform—satellite-based, UAV-based, airborne, or ground-based—the raw spectral images require a systematic series of operations to prepare the data for predictive modeling. Figure 2.9 conceptually illustrates a typical satellite image preprocessing workflow, transitioning from unprocessed imagery to spatially aligned and cleaned datasets usable for environmental applications. It serves as a foundational step before applying any ML model.

The steps to convert raw spectral images into spectral reflectance values are as follows [5, 155]:

1. ***Composition of Bands:*** The satellite imagery is composited into a single image in this initial stage by combining the tif files from each of the bands into a single tif file.

2. ***Mosaicing:*** Due to its broad coverage area, satellite imagery frequently arrives in numerous tiles. Mosaicing entails stitching these individual tiles to form a smooth and complete representation of the study area.

3. ***Clipping:*** A clipping technique aligns the satellite picture with the specified study site. Trimming the larger mosaicked image to the borders of the study region ensures the preservation of only pertinent data for future research.

4. ***Selection by location:*** The final step is to choose the areas of the clipped satellite imagery that precisely correspond to the sites covered by the ground truth dataset. This procedure ensures that the satellite images and the ground truth dataset are spatially aligned, allowing for the reliable comparison and integration of ground-level and remote sensing data.

In addition to the general preprocessing workflow, GIS tools play a vital role in preparing raw spectral images for analysis [156]. Platforms like QGIS, ENVI, and ArcMap provide comprehensive functionalities that facilitate each image preprocessing step.

1. ***QGIS [157]:*** It is an open-source GIS software that enables users to perform band stacking or composition, merging multiple spectral bands into a single, multi-layered image. QGIS also supports mosaicing to seamlessly stitch multiple image tiles, clipping to trim images to the boundaries of the study area, and spatial extraction to align satellite data with ground truth points.

2. ***ENVI [158]:*** It is a commercial software specialized for remote sensing, offering advanced tools for radiometric and geometric correction, noise reduction, and normalization, in addition to band composition and subsetting for regions of interest. ENVI also provides spectral analysis tools, such as vegetation indices and spectral profiling, which can be integrated into machine learning workflows.

3. ***ArcMap/ArcGIS [19]:*** It supports mosaicing, raster processing, clipping, and georeferencing tools to ensure accurate spatial alignment between imagery

and ground truth datasets. ArcGIS further enables seamless integration with external analytical tools, preparing the processed images for machine learning and predictive modeling.

In chapter 4, this thesis applies the ArcMap/ArcGIS [19] GIS tool to preprocess Landsat-8 [159] multispectral imageries of the LUCAS-2015 [2] and Rupnagar dataset [5] to propose the HyPLS-GPR hybrid ML model for Research Objective 3. It ensures that spectral images are accurate, spatially aligned, and structured for quantitative analysis, thereby forming a reliable foundation for downstream applications, including environmental monitoring [160].

2.5.2 Noise Removal

Noise in spectral data can arise from various sources, including atmospheric interference, sensor errors, and environmental factors [161]. Effective noise removal enhances data quality by eliminating unwanted signals while preserving important spectral features. The smoothing techniques applied for noise removal are as follows:

1. ***Savitzky-Golay Smoothing:*** This polynomial smoothing filter is widely used to reduce high-frequency noise in spectral data [162]. Fitting a polynomial to successive data points preserves the shape of absorption features, making it particularly useful for hyperspectral data [18, 163–166]. It is especially beneficial in retaining fine spectral details and improving model interpretability. The Savitzky–Golay filter has two main variants: the traditional integer-order filter and the more advanced non-integer (fractional-order) filter. While the integer-order filter uses whole-number derivatives and works effectively for general smoothing, the fractional-order version offers finer control over smoothing by allowing non-integer derivatives (e.g., 0.5 or 1.3). It improves the enhancement of subtle absorption features and has been shown to provide better prediction accuracy [163]. This technique can be applied using the Python library *scipy.signal.savgol_filter*[1]. A polynomial of degree x with a window length of y can be fitted for each pixel-wise spectrum. It preserves subtle absorption features

[1] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol_filter.html`

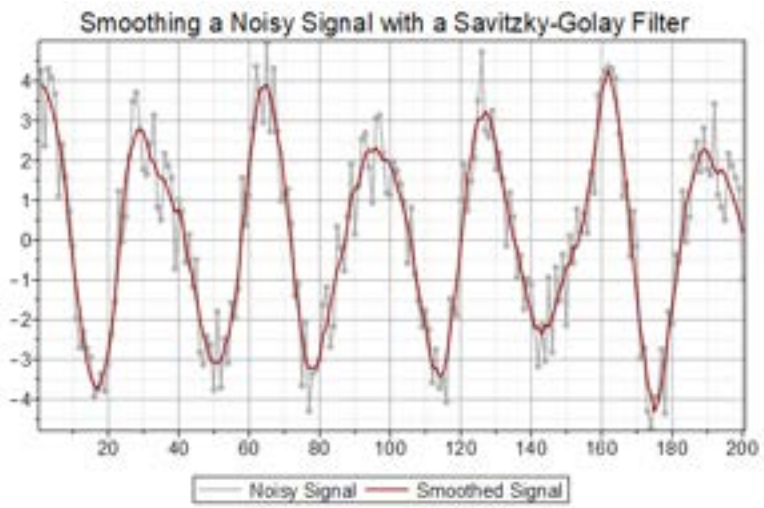

Figure 2.10: Savitzky-Golay smoothing

while effectively smoothing high-frequency noise. Savitzky-Golay smoothening is helpful before applying feature selection algorithms to retain informative band patterns. Figure 2.10 shows the effect of a Savitzky-Golay filter on a noisy signal. The grey curve represents the original noisy spectral data, while the red curve shows the smoothed output. The filter preserves the overall shape and peak positions of the spectral signal while reducing high-frequency noise.

2. ***Whittaker Smoothing:*** This method uses a penalized least squares algorithm to balance smoothness and data fidelity, effectively reducing noise while maintaining the integrity of spectral features [167]. Whittaker smoothing is suitable for datasets with irregular or missing data points, providing robust results. It is used via a custom implementation based on penalized least squares [168–170]. The regularization parameter (lambda) can be tuned using a subset of the data and the validation performance. It handles inconsistent patterns and effectively uses spectra with small gaps or spikes. Figure 2.11 demonstrates the Whittaker smoothing to spectral data. The left panel shows the original spectral band with significant noise, while the right panel displays the smoothed version after Whittaker filtering. This method effectively reduces random fluctuations and enhances the visibility of spectral trends.

3. ***Fourier Transform:*** It decomposes the spectral data into its frequency components, allowing the removal of noise by filtering out high-frequency

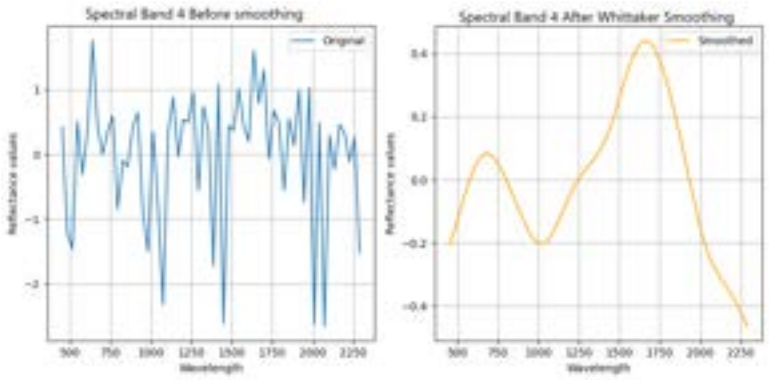

Figure 2.11: Whittaker smoothing

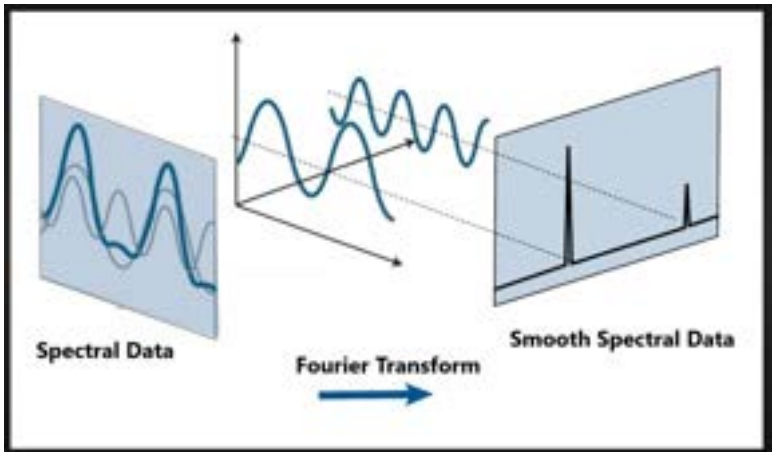

Figure 2.12: Fourier Transform smoothing

components while retaining the essential signal [166, 166]. It eliminates periodic or random noise and is widely applied in hyperspectral analysis. Each reflectance spectrum is transformed using Fast Fourier Transform (FFT) (*numpy.fft* [2]). High-frequency components (those above a defined threshold, determined by energy spectrum analysis) are filtered out. The inverse FFT is then applied to reconstruct the denoised signal, which helps remove periodic sensor noise. Figure 2.12 illustrates smoothing spectral data using a Fourier transform. The original spectral signal contains fluctuations and noise, which are decomposed into frequency components. By filtering high-frequency noise and reconstructing the signal, the resulting smooth spectral data highlights the main spectral features while minimizing minor variations.

[2] `https://numpy.org/doc/stable/reference/generated/numpy.fft.fft.html`

2.5.2.1 Review and Discussion on Noise Removal Techniques

All the noise removal techniques — Savitzky–Golay (SG), Fourier Transform (FT), and Whittaker smoother — have been widely applied for spectral preprocessing. However, the performance varies depending on the application and data characteristics. Zhang et al. [164] applied a fractional-order SG filter to Vis–NIR spectra for predicting soil phosphorus and achieved an R^2 of 0.70, outperforming the integer-order SG filter by better reducing baseline shifts. In a similar study by Zhang et al. [163], a non-integer SG filter approach applied to pig manure spectral data showed an R^2 of 0.72 for phosphorus prediction, indicating its robustness in capturing subtle spectral variations linked to nutrient content.

In contrast, the Fourier-Transform method, such as the Quantum Fourier Transform, was used by Anand et al. [171] for soil nutrient prediction using hyperspectral data. This approach achieved an R^2 of 0.178 for phosphorus, demonstrating its utility in high-dimensional spectral feature extraction. Another study by Wang et al. [166] utilized Fourier-Transform Infrared (FTIR) spectroscopy to predict soil potassium, achieving its best results with an R^2 value of 0.49.

Meanwhile, Whittaker smoothing has been proven effective for denoising and reconstructing scattered or incomplete spectral data. Krylov et al. [172] used it to recover fluorescence spectra obscured by scattering and reported lower RMSE than spline and polynomial-based methods. Another study by Liang et al. [170] applied Whittaker smoothing with enhanced gap-filling to reconstruct NDVI from multi-source imagery. The method achieved higher accuracy (R^2 = 0.836–0.926, RMSE = 0.04) than the SG filter.

Thus, while Savitzky–Golay excels in preserving peak shapes, Whittaker is advantageous for robust smoothing in noisy conditions, and Fourier-Transform-based techniques are well-suited for handling complex, high-dimensional spectral signals. Choosing the appropriate method depends on the noise characteristics, signal complexity, and the specific objectives of spectral analysis.

2.5.3 Spectral Normalization

Spectral normalization ensures consistency across different spectra. Proper normalization ensures that all spectral values are within a uniform scale, preventing the dominance of features with larger magnitudes [173]. The following techniques can be applied for data normalization:

1. ***Standard Normal Variate (SNV):*** It corrects spectral variations caused by factors like particle size, moisture content, and surface roughness. Centering and scaling each spectrum minimizes scattering effects and enhances model performance. It is beneficial for reducing systematic biases and ensuring prediction. SNV preprocessing is applied per-spectrum across all samples [174–177]. Specifically, for each spectrum, the mean and standard deviation are computed. Subsequently, each spectral point is centered by subtracting the mean and scaled by dividing by the standard deviation. It helps to mitigate the impact of variation and surface roughness effects, especially in field-acquired ground spectra.

2. ***Standard Scaler:*** It transforms spectral data to have a mean of zero and a standard deviation of one [178–181]. This method ensures that spectral features with different magnitudes are treated equally during model training. Standard scaling is essential when applying machine learning algorithms sensitive to feature scales [182]. This preprocessing technique uses the *StandardScaler*[3] module from scikit-learn. In this approach, scaling is applied globally across all samples and wavelengths, such that each spectral feature is standardized based on the dataset's overall mean and standard deviation. This normalization is crucial before training models that are sensitive to feature scales.

2.5.4 Feature Selection

Feature selection involves identifying the most relevant spectral bands for model development and training. Reducing dimensionality enhances model efficiency, reduces computational cost, and minimizes the risk of overfitting [33]. Selecting informative features also improves model interpretability, providing clearer insights

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

into the relationship between spectral data and predicting concentrations [152]. Various approaches are commonly employed for this purpose. *Recursive Feature Elimination (RFE)* iteratively removes the least essential features based on model-derived importance scores, refining the subset to enhance predictive accuracy [183]. *Principal Component Analysis (PCA)* reduces dimensionality by transforming correlated spectral bands into uncorrelated components, retaining most of the variance while simplifying the dataset [184]. *Successive Projections Algorithm (SPA)* selects minimally collinear features by maximizing vector projections, ensuring low redundancy among highly correlated hyperspectral bands [185]. Linear relationships between spectral bands and target variables can be captured using *Pearson Correlation Coefficient (PCC)*, which retains features with strong correlations [186]. Regression-based selection methods, such as *LASSO*, apply L1 regularization to eliminate less informative features by forcing coefficients to zero [187]. For large datasets and nonlinear dependencies, *XGBoost Feature Importance* ranks features based on the contribution to the predictive accuracy of gradient-boosted decision trees [188]. Together, these techniques offer complementary strategies for reducing dimensionality, enhancing model interpretability, and ensuring robust performance in spectral data analysis. A detailed discussion on feature selection techniques is presented in Chapter 3.

2.6 Summary

This chapter provides the technical background of spectral data in remote sensing, starting with the concept of remote sensing, its various types, and its applications in Earth observation. It introduces the electromagnetic spectrum and its components (UV, VIS, NIR, SWIR, MIR), which form the basis for understanding spectral measurements. The chapter explains key spectral concepts, including spectral bands, reflectance, and spectral signatures, which are crucial for interpreting surface materials. It then describes spectral imaging and outlines how spectral data is acquired using different platforms, including satellite, airborne, UAV, and ground-based sensors. It classifies spectral data into MSD and HSD, emphasizing the key characteristics and differences in resolution, data richness, and application scope. It also discusses the advantages of spectral data, followed by common challenges such as noise, redundancy, and environmental

variability. Finally, the chapter introduces an automated spectral data preprocessing pipeline, which includes converting raw spectral images into spectral reflectance value data, removing noise, normalizing the spectrum, and selecting features. These steps are crucial for enhancing prediction accuracy, reducing dimensionality, and facilitating effective ML-based modeling.

Chapter 3

Literature Review

This chapter discusses the various open spectral soil datasets that form the foundation for spectral-based research on soil nutrient prediction. It also shows the key performance metrics commonly used to evaluate spectral-based models for predicting soil nutrients. These metrics play a crucial role in evaluating the accuracy, robustness, and generalizability of machine learning models, serving as a foundation for comparing the effectiveness of different analytical approaches. Following this, the chapter provides a comprehensive review of feature selection techniques, machine learning, deep learning, and spectral indices for application to soil nutrient analysis using spectral data. It evaluates the strengths and limitations of existing studies, emphasizing the role of these methods in improving prediction accuracy and computational efficiency. The review also identifies critical gaps in current research and offers insights for future development of more scalable, interpretable, and application-oriented models to advance sustainable agricultural practices.

3.1 Open Spectral Soil Datasets

This thesis presents various openly accessible spectral soil nutrient datasets. These datasets provide diverse information on soil nutrients collected across extensive geographical regions, utilizing standardized soil sampling and laboratory analysis protocols. The data was carefully curated, encompassing various soil types and environmental conditions, enabling researchers to develop, train, and validate robust soil nutrient prediction models. The unique characteristics of each dataset emphasize

the need for models to generalize across diverse environmental and soil contexts.

Table 3.1 summarizes key open-access datasets including LUCAS-2009 [1], LUCAS-2015 [2], AfSIS [3], HYPERVIEW [4], Rupnagar dataset [5] and Univalle dataset [6]. These datasets cover various geographic regions, capturing diverse soil types and environmental conditions. Soil samples are typically collected from topsoil (0–20 cm), with some datasets including subsoil, and analyzed for key physicochemical properties. Spectral data are acquired using various tools, ranging from laboratory-grade spectroradiometers (as in LUCAS-2009 [1], AfSIS [3], and Univalle [6, 18]) to high-resolution satellite hyperspectral imaging systems (HYPERVIEW [4]), and include Landsat-8 [159] satellite multispectral systems (LUCAS-2015 [2] and Rupnagar dataset [5]). It includes spectral measurements covering the visible (400–700 nm), near-infrared (700–1100 nm), and short-wave infrared (1100–2500 nm) regions. Dataset sizes vary considerably: LUCAS-2009 [1] and LUCAS-2015 [2] are relatively large, containing thousands of samples across Europe, AfSIS [3] includes several thousand samples from sub-Saharan Africa, while HYPERVIEW [4] includes several thousand samples from Poland. Datasets like Rupnagar [5], and Univalle [6, 18] are smaller, comprising a few hundred samples, which are suitable for local or regional studies. These datasets provide comprehensive resources for calibrating, validating, and developing machine learning and remote sensing models to predict soil nutrients.

This thesis uses all open datasets to benchmark spectral-based soil nutrient prediction using ML techniques under varying spectral ranges, sensors used, soil samples, and spectral bands. Table 3.2 presents a concise mapping of the open-access soil spectral datasets used in this work, indicating the spectral type, the soil nutrients predicted, and the research objectives addressed with each dataset. It also shows whether feature selection, deep learning, or spectral indices are applied for each dataset. This table highlights the diversity of datasets in terms of spectral range, nutrient coverage, and analytical approaches, demonstrating how each contributes to the predictive analyses presented in this thesis.

Table 3.1: Open soil spectral datasets available for soil nutrient prediction

Dataset and Year	Region	HS/MS	Soil nutrients	No. of soil samples	No. of spectral bands	Spectral range (nm)	Acquisition tool
LUCAS-2009 [1]	23 European countries	HS	N, P, K, SOC, pH in H2O, pH in Cacl2, CEC, CaCO3	19036	4200	400-2500	Ground Based: FOSS Spectroscope
LUCAS-2015 [2]	28 European countries	MS	N, P, K, SOC, pH in H2O, pH in Cacl2, EC, CaCO3	21859	7	450-2290	Spaceborne: Landsat-8
AfSIS (2009) [3]	19 Sub-Saharan African countries	HS	P, pH, SOC, Ca	1157	3578	1333-16673	Ground Based: Spectrometer
HYPERVIEW (2020) [4]	Poland	HS	P, K, Mg, pH	1732	150	462-942	Spaceborne: Intuition-1 satellite
Rupnagar dataset (2020) [5]	India	MS	pH, SOC	60	11	450-2290	Spaceborne: Landsat-8
Univalle dataset (2022) [6, 18]	Colombia	HS	P	152	3501	200-900	Ground Based: UV–VIS JASCO V-750 Spectrophotometer

HS: Hyperspectral, **MS:** Multispectral

Table 3.2: Mapping of open-access soil spectral datasets with research objectives

Dataset	HS/MS	Soil nutrients predicted in this thesis	Feature Selection (RO2)	Deep Learning (RO3)	Spectral Indices (RO4)
LUCAS-2009 [1]	HS	pH, SOC, P, Ca	P.S.: × C.S.: ✓(in RO2)	P.S.: Limited [65, 189–191] C.S.: ✓(in RO3)	P.S.: × C.S.: ✓(in RO4)
LUCAS-2015 [2]	MS	N, P, K, SOC, pH in H2O, pH in Cacl2, EC, CaCO3	N/A (few spectral bands)	N/A (small dataset size)	P.S.: × C.S.: × (HS focus)
AfSIS [3]	HS	pH, SOC, P, Ca	P.S.: × C.S.: ✓(in RO2)	P.S.: Limited [65, 189, 190] C.S.: ✓(in RO3)	P.S.: × C.S.: ✓(in RO4)
HYPERVIEW [4]	HS	P, K, Mg, pH	P.S.: × C.S.: ×	P.S.: Limited [17, 192] C.S.: ✓(in RO3)	P.S.: × C.S.: × (DL focus)
Rupnagar dataset [5]	MS	pH, SOC	N/A (few spectral bands)	P.S.: Limited [5] (basic MLP used) C.S.: N/A (small dataset size)	P.S.: × C.S.: × (HS focus)
Univalle dataset [6, 18]	HS	P	P.S.: × C.S.: ×	P.S.: Limited [18] C.S.: ✓(in RO3)	P.S.: × C.S.: × (DL focus)

Note: PS: Previous status, CS: Current status; ✓: Performed, ×: Not performed, N/A: Not applicable

3.1.1 A Database of Land Use and Coverage Area frame Survey (LUCAS-2009) [1]

The Land Use and Coverage Area Frame Survey (LUCAS), launched in 2001 by Eurostat, is designed to monitor land cover and land use across the European Union using a harmonized, standardized methodology. Figure 3.1 illustrates the spatial distribution of LUCAS-2009 [1]. The map displays the locations of the samples,

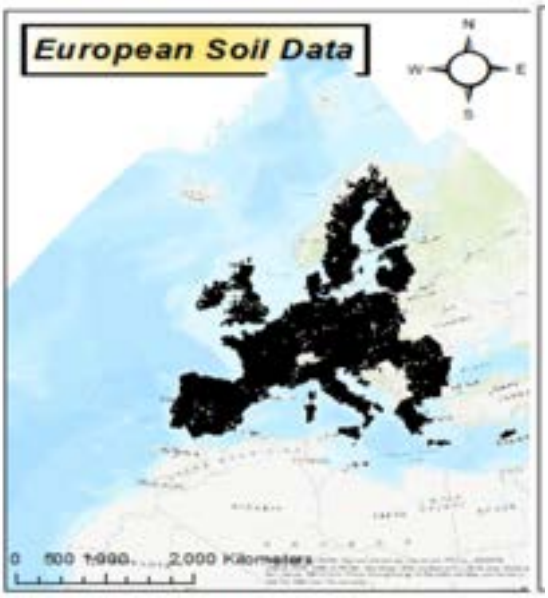

Figure 3.1: Spatial distribution of LUCAS-2009 [1]

Table 3.3: Summary statistics for LUCAS-2009 soil nutrients

Property	Min	Max	Mean	Std Dev	Skewness
SOC ($g.kg^{-1}$)	0.00	586.80	50.0027	91.3070	3.6751
N ($g.kg^{-1}$)	0.00	38.60	2.9248	3.7551	3.7628
P ($mg.kg^{-1}$)	0.00	1366.40	30.0759	32.8550	6.3375
K ($mg.kg^{-1}$)	0.00	7342.00	197.0519	229.2987	8.4694
pH_H_2O	3.21	10.08	6.1984	1.3530	-0.0784
pH_$CaCl_2$	2.57	9.25	5.5944	1.4260	-0.1609
$CaCO_3$ ($g.kg^{-1}$)	0.00	944.00	51.6014	125.3180	2.9822
CEC ($cmol.kg^{-1}$)	0.00	234.00	15.7568	14.4847	4.2440

with each data point in this figure representing a sample. It also provides context for the variability in soil nutrients captured in the dataset, which is essential for modeling and analysis. Table 3.3 presents the descriptive statistics of the LUCAS-2009 soil dataset [1] used in this thesis. The wide variation and positive skewness observed in most properties indicate heterogeneous soil fertility and compositional diversity across sampling sites, highlighting the need for robust predictive modeling.

The following are the salient features of the LUCAS-2009 dataset [1].

1. ***Grid Size and Sampling Strategy:*** Soil sampling across the European Union was done using a systematic grid-based approach. A regular 2 km × 2 km grid was overlaid across the region, and sampling points were selected at the grid

intersections to ensure uniform spatial coverage throughout the area. During the 2009 campaign, approximately 21,000 sampling locations were identified across 23 EU Member States.

2. ***Soil Depth and Sample Collection:*** The sampling protocol targeted the top 0–30 cm of the soil profile, which is most influenced by agricultural practices and root activity. This layer is crucial for assessing nutrient availability and other indicators of soil health. Standardized sampling procedures were followed to ensure consistency across all regions and minimize sampling bias.

3. ***Physicochemical Property Analysis:*** The collected soil samples were analyzed in certified laboratories for key physicochemical properties. These include pH measured in both H_2O and $CaCl_2$, Organic Carbon (OC), total Nitrogen (N), Phosphorus (P), Potassium (K), Cation Exchange Capacity (CEC), and Calcium Carbonate ($CaCO_3$). These parameters are essential for understanding soil fertility, nutrient availability, and quality.

4. ***Spectral Data Acquisition:*** Alongside laboratory analysis, each soil sample underwent spectral scanning in the VNIR range of 400–2500 nm using FOSS Spectroscope [193]. The VNIR spectra provide rich information on organic and mineral components in soil and are widely used in predictive modeling using machine learning.

5. ***Data Format and Preprocessing:*** The released dataset primarily contains raw spectral reflectance values with minimal correction, comprising 19,036 soil samples with 4,200 spectral input features. Researchers are expected to perform additional preprocessing steps, such as spectral smoothing and data normalization, depending on the modeling requirements and objectives of the study.

3.1.2 LUCAS-2015 Topsoil Dataset with Landsat-8 Imagery [2]

The 2015 edition of the LUCAS, coordinated by Eurostat, expands upon the previous LUCAS surveys by incorporating extensive topsoil measurements and remote sensing



Figure 3.2: Spatial distribution of LUCAS-2015 [2]

data. The LUCAS 2015 [2] topsoil campaign was conducted across 28 European countries. This comprehensive effort supports land use and soil nutrient assessments at a continental scale. This thesis integrates the LUCAS-2015 [2] topsoil dataset with Landsat 8-9 OLI/TIRS C2 Level 2 [159] multispectral imagery, obtained from the USGS EarthExplorer portal [1], to investigate soil nutrient distributions and land use patterns. Figure 3.2 illustrates the spatial distribution of LUCAS-2015 data. The map displays the locations of the samples, with each data point in this figure representing a sample. It provides an overview of the geographic coverage and helps assess the dataset's representativeness for soil nutrient analysis across different regions. Table 3.4 presents the summary statistics of the LUCAS-2015 soil dataset [2]. It reports the minimum, maximum, mean, standard deviation, and skewness for each property, reflecting the variability and heterogeneity in soil composition across sampling sites. Positive skewness in most nutrient parameters indicates a distribution with a long right tail, whereas pH values are approximately symmetric, highlighting the overall diversity of soil chemical properties.

[1] https://earthexplorer.usgs.gov/

Table 3.4: Summary statistics for LUCAS-2015 soil nutrients

Property	Min	Max	Mean	Std Dev	Skewness
SOC ($g.kg^{-1}$)	0.10	560.20	43.2760	76.6967	4.3601
N ($g.kg^{-1}$)	0.00	38.50	3.1047	3.6720	3.9049
P ($mg.kg^{-1}$)	0.00	1017.60	32.6138	32.3722	4.1454
K ($mg.kg^{-1}$)	0.00	10030.90	195.5297	232.3838	13.4384
pH_H_2O	3.17	10.37	6.1308	1.3453	0.0059
pH_$CaCl_2$	2.60	10.00	5.7452	1.3984	-0.1463
$CaCO_3$ ($g.kg^{-1}$)	0.00	976.00	57.2682	135.3005	2.9117
CEC ($cmol.kg^{-1}$)	0.31	969.00	26.0915	33.3396	6.0070

The following are the salient features of the LUCAS-2015 dataset [2].

1. **Grid Size and Sampling Strategy:** Sampling points were geo-referenced and distributed across a harmonized EU-wide grid. Approximately 21,859 topsoil samples were collected from 28 Member States of the European Union, each representing the 0–30 cm soil layer to ensure consistency with earlier LUCAS surveys.

2. **Soil Depth and Sample Collection:** The 0–30 cm topsoil layer was sampled uniformly across all locations following standardized protocols to ensure comparability and reduce sampling bias.

3. **Physicochemical Property Analysis:** Samples were analyzed for key properties including pH (in H_2O and $CaCl_2$), Organic Carbon (OC), Nitrogen (N), Phosphorus (P), Potassium (K), Calcium Carbonate ($CaCO_3$), and Electrical Conductivity (EC). Soil texture and other agronomic attributes were also recorded.

4. **Spectral Data Acquisition:** Landsat 8–9 OLI/TIRS [159] multispectral imagery from 2015 was used to extract seven surface spectral features related to soil nutrients. With a 30 m spatial resolution, the imagery covers multiple VNIR bands, making it suitable for land cover and soil analysis.

5. **Data Format and Preprocessing:** Landsat Level-2 imagery was obtained from the USGS EarthExplorer portal [2]. Although pre-corrected radiometrically

[2] https://earthexplorer.usgs.gov/

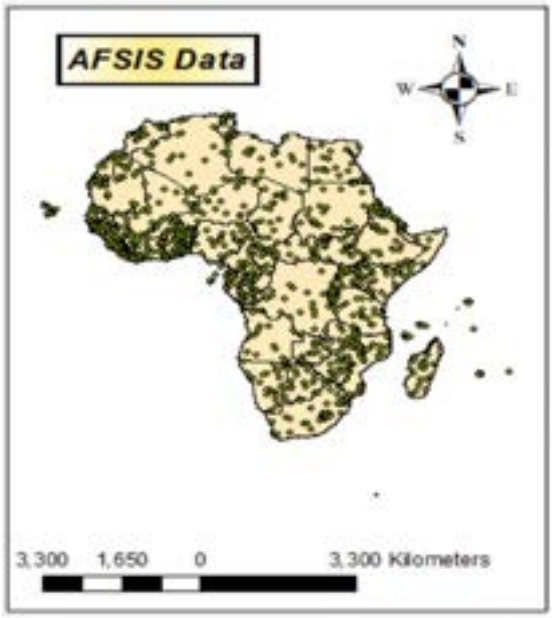

Figure 3.3: Spatial distribution of AfSIS dataset [3]

and atmospherically, additional preprocessing—such as noise removal and data normalization —is typically performed by researchers.

3.1.3 A Database of Africa Soil Information Service (AfSIS) [3]

The Africa Soil Information Service (AfSIS) database is a consolidated repository of soil profile data collected from across Sub-Saharan Africa. Figure 3.3 illustrates the spatial distribution of AfSIS [3]. The map displays the locations of the samples, with each data point in this figure representing a sample. This visualization highlights the geographic coverage of the dataset across Africa, providing insight into the spatial variability of soil nutrients captured for analysis and modeling. Table 3.5 presents the summary statistics of the soil nutrients in the AfSIS dataset [3]. The negative values reflect monotonically transformed data, as noted in the Kaggle repository [194]. Since the original ranges or inverse transformation functions are not provided, this thesis retains and evaluates models using these transformed values. Consequently, units for each soil nutrient are not reported.

The following are the salient features of the AfSIS dataset [3].

1. **Grid Size and Sampling Strategy:** The AfSIS [3] dataset is designed to support soil nutrient mapping at moderate spatial resolution, typically ranging from 1

Table 3.5: Summary statistics for AfSIS soil nutrients

Property	Min	Max	Mean	Std Dev	Skewness
SOC	-0.85	7.61	0.08	1.14	2.44
P	-0.41	13.26	-0.01	0.99	7.43
pH	-1.88	3.41	-0.02	0.92	0.86
Ca	-0.53	9.64	0.006	1.07	4.70

km^2 to 10 km^2, depending on the soil nutrient analyzed.

2. **Soil Depth and Sample Collection:** Samples were collected systematically from both topsoil and subsoil layers. The 0–20 cm topsoil represents the agriculturally active zone, while the 20–50 cm subsoil captures deeper nutrient pools. Standardized protocols ensured uniformity across all sampling locations.

3. **Physicochemical Property Analysis:** Measured soil nutrients include pH, Organic Carbon (OC), Phosphorus (P), and Calcium (Ca). These parameters provide essential information for assessing soil fertility, digital soil mapping, and precision agriculture applications.

4. **Spectral Data Acquisition:** Spectral reflectance data were recorded for each sample in the 1,333–16,673 nm range using high-resolution spectroradiometers. These spectral measurements capture the optical properties of soils and are crucial for machine learning-based predictions of soil nutrients.

5. **Data Format and Preprocessing:** The dataset is partially preprocessed for quality control but requires additional processing for modeling, including noise removal and data normalization. The dataset contains 1,157 soil samples, each with 3,578 spectral input features. Researchers typically perform further spectral preprocessing to enhance model performance and remove noise.

3.1.4 HYPERVIEW Hyperspectral Soil Dataset [4]

The HYPERVIEW dataset, introduced by the Hyperview Challenge organizers, contains hyperspectral imagery of soil patches collected from agricultural regions in Poland. It serves as a benchmark for predicting soil nutrients using satellite-based hyperspectral sensing. Figure 3.4 shows a raw hyperspectral sample alongside its

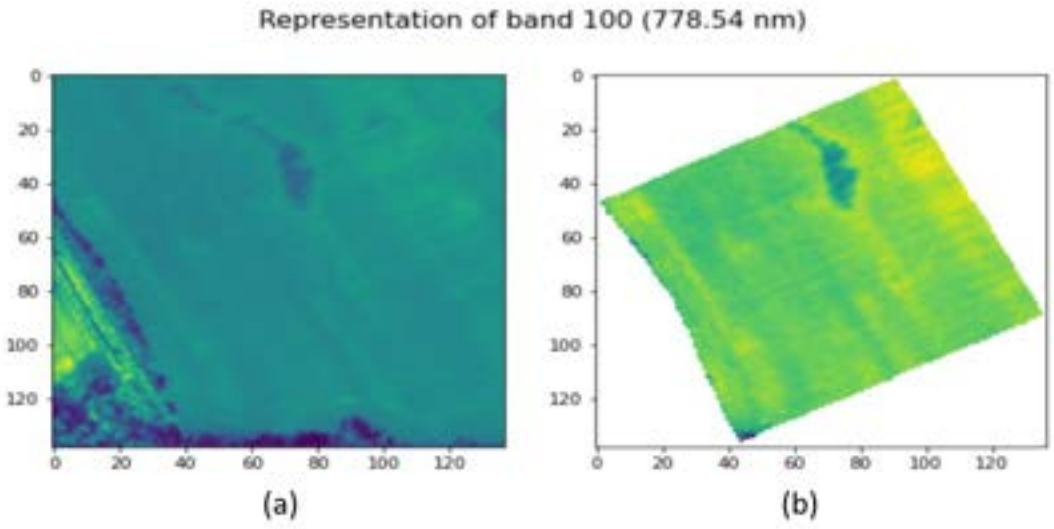

Figure 3.4: (a) HYPERVIEW dataset [4] instance illustrating an image of the raw hyperspectral band (b) the corresponding masked region of interest. [17]

masked region, focusing on the area designated for soil nutrient analysis. It highlights the preprocessing step to isolate relevant regions, ensuring that the spectral data accurately represents the soil nutrients of interest. Table 3.6 summarizes the key statistics of soil nutrients in the HYPERVIEW dataset [4]. These values do not include units because the data are preprocessed or normalized measurements provided for modeling purposes. The original measurement scales or conversion factors are not provided in the dataset documentation, so this thesis retains the values for evaluation. The HYPERVIEW dataset [4] includes a dataset-specific evaluation measure called the score, defined by the organizers of the Hyperview Challenge [195]. It compares a model's prediction error to the fixed RF+KNN baseline included with the dataset by taking the ratio of the two errors (MSE_{algo} and MSE_{bl}). This brief description is provided here since the score forms part of the dataset's official evaluation protocol. A lower score is preferable for predicting soil nutrients. Equation 3.1 shows the computation of the score metrics.

$$\text{Score} = \frac{\text{MSE}_{\text{algo}}}{\text{MSE}_{\text{bl}}} \tag{3.1}$$

The following are the salient features of the HYEPRVIEW dataset [4].

1. **Grid Size and Sampling Strategy:** The HYPERVIEW dataset [4] comprises

Table 3.6: Summary statistics of HYPERVIEW soil nutrients

Property	Min	Max	Mean	Std Dev	Skewness
P	20.3	325.0	70.30	29.50	2.08
K	21.1	625.0	227.99	61.89	1.57
Mg	26.8	400.0	159.28	39.87	0.73
pH	5.6	7.8	6.78	0.26	-0.01

2,886 hyperspectral image patches, each containing 150 spectral bands spanning the 462–942 nm range. Spectral imagery was acquired using the Intuition-1 satellite [196] sensor from August to November 2020. The data were stored in `.npz` format, enabling compact and structured access to hyperspectral values.

2. **Soil Depth and Sample Collection:** Each image patch corresponds to a specific ground location. Soil samples were collected to ensure spatial correspondence with the satellite imagery, enabling the annotation of spectral signatures with ground-truth soil nutrients.

3. **Physicochemical Property Analysis:** A total of 1,732 samples include measured soil nutrients: Phosphorus (P), Potassium (K), Magnesium (Mg), and soil pH. These measurements provide critical information for estimating nutrients and training models. The remaining 1,154 samples were unlabeled and reserved for benchmarking and model evaluation.

4. **Spectral Data Acquisition:** Each sample includes a mask to isolate the region of interest, ensuring that only soil pixels are considered during analysis. Hyperspectral reflectance data capture fine-grained spectral variations, enabling advanced modeling of soil nutrients using machine learning and deep learning techniques.

5. **Data Format and Preprocessing:** Depending on the model requirements, the dataset is partially preprocessed for quality control but may require additional steps, such as normalization or spectral smoothing.

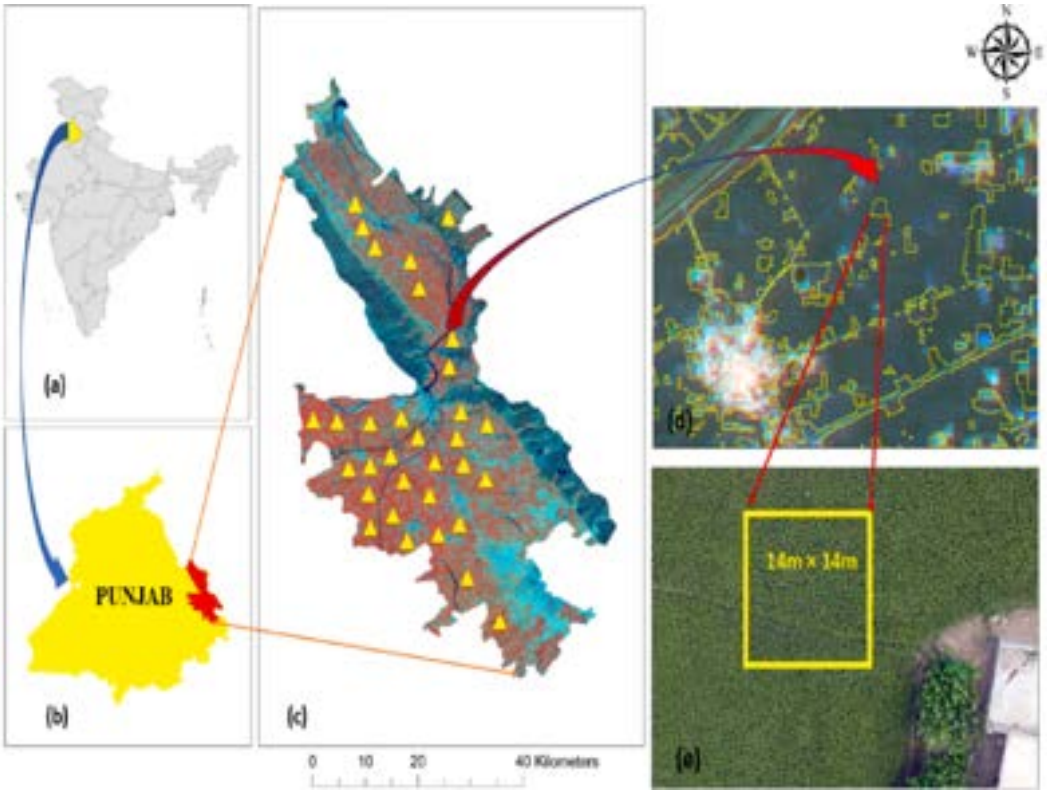

Figure 3.5: Spatial distribution of Rupnagar region [5]

Table 3.7: Summary statistics of Rupnagar dataset soil nutrients

Property	Min	Max	Mean	Std Dev	Skewness
SOC (in %)	0.150	0.900	0.553	0.163	-0.289
pH	6.550	8.040	7.211	0.456	0.315

3.1.5 Rupnagar Soil Dataset with Landsat-8 Imagery [5]

This dataset focuses on Rupnagar district in Punjab, India, located at approximately 30.97°N latitude and 76.53°E longitude, covering an area of 1,440 km^2. 60 soil samples were collected from the region and analyzed at the Indian Institute of Sugarcane Research (IISR), Lucknow. The dataset includes SOC and soil pH, covering a diverse range of soil conditions from acidic to alkaline. Figure 3.5 illustrates the study area in a series of maps: (a) Map of India highlighting the state of Punjab, (b) Map of Punjab with Rupnagar District distinctly marked in red, (c) standard false colour composite image of Rupnagar District, (d) Agricultural zones within the study region where wheat is cultivated, with all other Land Use/Land Cover categories masked out, and (e) A representative wheat field showing the pixel area used for ground truth data collection. Table 3.7 presents the summary statistics of soil nutrients in the Rupnagar dataset,

including SOC and pH. SOC values are expressed as percentages, while pH is unitless because it represents a logarithmic scale of acidity, not a direct measurement with physical units. These statistics provide an overview of the dataset's variability and distribution of the soil nutrients.

The following are the salient features of the Rupnagar dataset [5].

1. **Grid Size and Sampling Strategy:** Field sampling was conducted across representative sites in the Rupnagar district to ensure coverage of diverse soil conditions covering an area of 1,440 km^2. Soil samples were collected using standardized protocols, and laboratory analysis provided pH values that reflected spatial variability in the region.

2. **Soil Depth and Sample Collection:** Soil samples were collected from the topsoil layer, targeting the agriculturally active zone. Sampling locations were selected to capture heterogeneity in soil composition and land-use patterns across the district.

3. **Physicochemical Property Analysis:** The primary measured property in this dataset is soil pH. These values serve as a benchmark for assessing the relationship between multispectral reflectance and soil acidity.

4. **Spectral Data Acquisition:** Landsat-8 [159] level-1 multispectral imagery from 2020 was used, obtained from the USGS Earth Explorer portal [3]. Two tiles covering the study area were included, encompassing 11 spectral bands ranging from visible to thermal infrared (450–2290 nm). Key bands include: Band 1 (Coastal/Aerosol), Band 2 (Blue), Band 3 (Green), Band 4 (Red), Band 5 (Near Infrared), Band 6 (SWIR 1), Band 7 (SWIR 2), Band 8 (Panchromatic), Band 9 (Cirrus), Band 10 (Thermal Infrared 1), and Band 11 (Thermal Infrared 2), all with a spatial resolution of 30 meters.

5. **Data Format and Preprocessing:** The Landsat imagery is radiometrically and atmospherically corrected (Level 1), but additional preprocessing, such as noise removal or normalization, is typically performed to improve the accuracy of soil nutrient modeling.

[3] `https://earthexplorer.usgs.gov/`

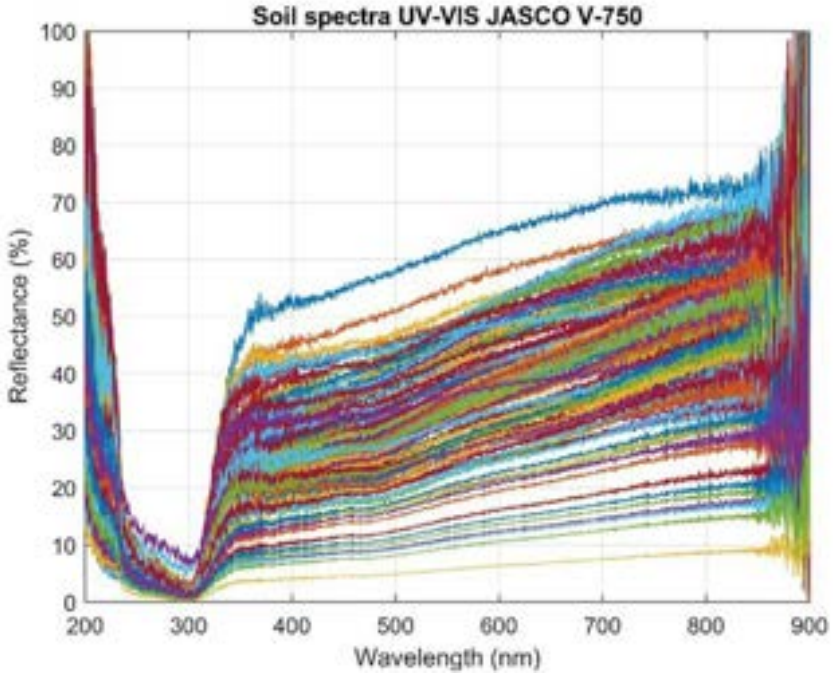

Figure 3.6: Univalle database of reflectance spectra of 152 soil samples [18].

Table 3.8: Summary statistics of phosphorus content in Univalle dataset

Statistic	**Phosphorus (mg/kg^{-1})**
Min	400.30
Max	212571.00
Mean	109052.32
Std Dev	64700.89
Skewness	0.043

3.1.6 Univalle Hyperspectral Soil Phosphorus Dataset [6]

The Univalle dataset [6, 18] contains laboratory-acquired hyperspectral reflectance data of soil samples from Santander de Quilichao, Colombia. It was developed to support machine learning-based prediction of total phosphorus concentration in soil [18]. The dataset emphasizes controlled soil-fertilizer mixtures, providing a foundation for building data-driven models in precision agriculture. Figure 3.6 shows the UV-VIS reflectance spectra of 152 soil samples measured using the JASCO V-750 spectrophotometer over the 200–900 nm wavelength range. The spectra exhibit diverse reflectance patterns across samples, highlighting variations in soil composition. Table 3.8 summarizes the key statistics of phosphorus content in the Univalle dataset [6, 18], reporting the minimum, maximum, mean, standard deviation, and skewness.

The values are expressed in milligrams per kilogram (mg/kg), providing an overview of the distribution and variability of phosphorus in the soil samples.

The following are the salient features of the Univalle dataset [6].

1. **Grid Size and Sampling Strategy:** The Univalle dataset comprises 152 soil samples derived from a single location in Santander de Quilichao, Colombia ($3°06'01.1''$N, $76°30'59.9''$W), characterized by acidic clay soil. Instead of spatial grid sampling, the dataset was generated under controlled laboratory conditions by systematically mixing soil with diammonium phosphate (P_2O_5) fertilizer in varying proportions.

2. **Soil Depth and Sample Collection:** All samples were obtained from a single, well-characterized soil type native to Santander de Quilichao, Colombia. The soil exhibits a high clay content (64.8%), an acidic pH (4.3), and distinct physicochemical properties. This homogeneity ensures consistency in spectral measurements.

3. **Physicochemical Property Analysis:** Laboratory analysis was performed on all 152 samples to determine total phosphorus content using standardized chemical methods. This ground truth was used for training and evaluating phosphorus prediction models.

4. **Spectral Data Acquisition:** Spectral reflectance data were acquired using a high-resolution JASCO V-750 UV–VIS spectrophotometer in the 200–900 nm range, capturing 3,501 spectral bands per sample. The equipment utilizes a symmetrical double-beam system with halogen and deuterium light sources for precise reflectance measurements.

5. **Data Format and Preprocessing:** Spectral data were initially stored in CSV files and later structured into MATLAB `.mat` format. To remove noise—especially prevalent at 200–250 nm and 860–900 nm—a Savitzky-Golay filter with a window size of 41 and a polynomial degree of 3 was applied.

3.2 Performance Metrics

The crucial metrics that help to assess the prediction accuracy of regression models are as follows [197]:

1. **Coefficient of Determination *(R^2):*** It represents the proportion of the variance in the dependent variable, which the linear regression model explains [198]. R^2 ranges from 1 to 0, where 1 indicates that the model explains all the variance, while 0 means it explains none of the variance. Equation 3.2 shows the computation of the coefficient of determination.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{3.2}$$

   where y_i are the observed values, $\hat{y}_i$ are the predicted values, $\bar{y}$ is the mean of observed values, and n is the number of observations.

2. ***Root Mean Square Error (RMSE):*** It measures the average size of the discrepancies between the actual observed values and the projected values [199]. Better model performance is indicated by lower RMSE values, with a value of 0 signifying flawless predictions. Equation 3.3 shows the computation of RMSE.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.3}$$

3. ***Mean Squared Error (MSE):*** It measures the average squared difference between the predicted values and the actual (observed) values in a dataset. The lower the MSE, the better the model's predictions match the data [195]. Equation 3.4 shows the computation of MSE.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.4}$$

4. ***Mean Absolute Error (MAE):*** It represents the average error between predictions and the true values. Expressed on the same scale as the dependent variable [200].

Equation 3.5 shows the computation of MAE.

$$\frac{1}{N}\sum_{i=1}^{n}(y_i - \bar{y_i}) \tag{3.5}$$

5. ***Ratio of Performance to Deviation (RPD):*** It evaluates the ratio between the standard deviation and root mean square error. An increased RPD indicates improved model performance [201]. A decent model has an RPD value above 2, whereas outstanding models typically have values above 3. Equation 3.6 shows the computation of RPD.

$$RPD = \frac{Standard\,Deviation}{RMSE} \tag{3.6}$$

3.3 Feature Selection Techniques for Soil Nutrient Prediction using Spectral Data

Feature selection is crucial for researchers to remove unnecessary features before making a prediction. The feature selection method aims to find the smallest subset of features from the original without reducing the accuracy rate [202]. This thesis explains specific feature selection techniques to predict soil nutrients. Table 3.9 presents some feature selection techniques employed by previous researchers for soil nutrient prediction, along with the limitations. It summarizes the commonly adopted methods and the effectiveness of these methods in enhancing model performance for predicting soil nutrients. The following is a list of some feature selection techniques used in previous studies for soil nutrient prediction:

1. ***Principal Component Analysis (PCA):*** It reduces data dimensionality by transforming correlated spectral bands into a set of linearly uncorrelated principal components [90, 91, 206]. PCA retains components that explain most of the variance, simplifying the model without significant loss of information. However, it is widely used for visualizing complex spectral data and improving model generalization. It is performed using a Python library

Table 3.9: Previous studies using feature selection techniques for soil nutrients prediction

Previous Work	HS/MS	Region	Nutrient	Spectral Data Tool	Feature Sel. / Absorp. Bands (nm)	Range (nm)	Model / Results	Limitation
Peng et al. (2024) [203]	HS	China	N, SOC	UAV: Gaia-Sorter HSI, Private soil data	Feature Sel.: Correlation Coefficient, SPA; Absorp. Bands: 1400, 1900, 2200	350–2500	RF, PLSR, RBF; R^2=0.97 (RF)	Only linear relations; noise sensitive; limited to pixel-level
Yuzugullu et al. (2024) [204]	MS	Europe	SOC	Spaceborne: Sentinel-2, Private soil data	Correlation Coefficient; Absorp. Bands not mentioned	443–1915	LGBM, RF, MLP; best R^2=0.54 (LGBM)	Linear relation only; single nutrient; noise sensitive
Shi et al. (2023) [153]	HS	China	SOM	Ground-Based: Spectrometer, Private soil data	CARS; Absorp. Bands: 400–800, 1600–2000, 2200–2400	400–2450	PLSR; RMSE=3.59	Noise sensitive; single nutrient
Zhang et al. (2023) [187]	HS	China	N	Spaceborne: ZY1-02D/A HSI, Private soil data	LASSO	400–2500	MLR, PLSR, SVM; R^2=0.62	Noise sensitive; single nutrient
Zhao et al. (2022) [190]	HS	Africa	SOC	Ground Based: Spectrometer; LUCAS-2009, AfSIS	CNN-based feature selection	599–7500	CNN; R^2=0.91	Only one nutrient targeted
Guo et al. (2022) [205]	HS	China	SOM	Spaceborne: ZY1-02D HSI, Private soil data	Correlation + LASSO; Absorp. Bands: 559, 567, 714, 731, 1964, 2048	400–2500	MLR, PLSR, RF; best R^2=0.92 (RF)	Linear relation; noise sensitive; single nutrient
Zhang et al. (2022) [90]	HS	China	Cr, Zn	Ground Based: Spectrometer, Private soil data	PCA	400–2500	Deep Forest (DF21); R^2=0.58	PCA introduces uncorrelated vars; noise increase
Guo et al. (2022) [33]	HS	China	pH	Ground Based: Spectrometer, Private soil data	SVM-RFE (20 features)	–	XGBoost; R^2=0.68, MAE=0.16, RMSE=0.26	Hard to interpret; single nutrient
Xu et al. (2022) [91]	HS	China	Fe	Ground Based: Spectrometer, Private soil data	PCA; Absorp. Bands: 986, 406, 1006, 978, 426	400–1010	RF, XGBoost, CatBoost, SVR; best RMSE=2.465 (SVR)	PCA uncorrelated vars; noise increase

Note: HS-Hyperspectral, MS-Multispectral

Table 3.9: Previous studies using feature selection techniques for soil nutrients prediction

Previous Work	HS/MS	Region	Nutrient	Spectral Data Tool	Feature Sel. and Absorp. Band (nm)	Spectral Range (nm)	Model and Results	Limitation
Wang et al. (2021) [93]	HS	China	Cr	Ground Based: Spectrometer, Private soil data	Discriminant Analysis; Absorp. Bands: 1004, 853, 2035	401–2400	DNN; Accuracy=96.25%	Linear assumption; limited validation
Pechanec et al. (2021) [206]	HS	Czech Republic	N	Ground Based: Spectrometer, Private soil data	PCA; Absorp. Band: 1300–3000	325–1075	PLSR; R^2=0.44	PCA noise increase; uncorrelated vars
Guo et al. (2021) [185]	HS	China	P, K	Ground Based: Spectrometer, Private soil data	SPA, CARS; Absorp. Bands: 1200, 1400, 1780, 2200	400–2450	PLSR, SVM; best R^2=0.75 (SPA-PLSR)	Noise sensitive; pixel-level limitation
Dong et al. (2021) [207]	HS	China	SOM	Ground Based: Spectrometer, Private soil data	CARS; Bands: 1400, 1900, 2200	350–2500	Cubist, PLSR, SVM, ANN; Best: Cubist (R^2=0.86)	Sensitive to noise; Examines only one nutrient.
Meng et al. (2021) [161]	HS	China	SOM	Spaceborne: Gaofen-5 satellite, Private soil data	RFE; Bands: 430–900, 1050–1350, 1451–1771, 1982–2450	390–2513	RF; R^2=0.84	Difficult to interpret; Single-nutrient focus.
Zhang et al. (2021) [208]	HS	China	SOM	Ground Based: Spectrometer, Private soil data	PCA; Bands: 400–780	400–2400	PLS-SVM; R^2=0.85	PCA uncorrelated data; Increased noise.
Guo et al. (2021) [185]	HS	China	SOM	Ground Based: Spectrometer, Private soil data	SPA, CARS; Bands: 1200, 1400, 1780, 2200	400–2450	CARS-SVM; R^2=0.90	Noise sensitivity; Pixel-level limitation.
Sabetizade et al. (2021) [209]	MS	Iran	SOC	Ground Based: Spectrometer, Private soil data	RFE; Bands: 950–1000, 1500–2000, 3600–3700, 5800–7500	400–2450	Cubist, PLSR; Best: Cubist (R^2=0.96)	Difficult interpretation; Single-nutrient study.
Zhu et al. (2020) [210]	HS	China	SOM	Ground Based: Spectrometer, Private soil data	PCA; Bands: 1400, 1900, 2200	400–2400	RF, Linear Regression; Best: RF (R^2=0.93)	PCA uncorrelated data; Noise increased.
Bangelesa et al. (2020) [211]	HS	Africa	SOC	Ground Based: Spectrometer, Private soil data	RFE; Bands: 400–700	350–2500	RF, PLSR; Best: RF (R^2=0.90)	Difficult interpretation; One nutrient only.
Zhang et al. (2019) [149]	MS	China	N	Ground Based: Spectrometer, Private soil data	Ant Colony Optimization; Bands: 943, 1004, 1097, 1351, 1550, 1710, 2123, 2254	800–2564	SVM; R^2=0.99	High computational cost; Single-nutrient study.
Zhang et al. (2018) [92]	Both	China	pH	Spaceborne: Landsat-8 + Ground: Spectrometer, Private soil data	PCA	350–1000	PLSR; RMSE=0.26	PCA data uncorrelated; Noise increase.
Peng et al. (2014) [212]	HS	China	SOC	Ground Based: Spectrometer, Private soil data	SPA; Bands: 800, 1000, 1400, 1900–2450	350–2500	SVM, PLSR; Best: SVM (R^2=0.73)	Pixel-level limitation; One nutrient only.

Note: HS-Hyperspectral, MS-Multispectral

sklearn.decomposition.PCA[4]. The number of components can be set to retain 95% of the total variance. A key limitation is that PCA causes a loss of interpretability because the transformed components no longer correspond to the original spectral bands.

2. ***Recursive Feature Elimination (RFE):*** It is an iterative technique that trains a model, evaluates feature importance, and removes the least significant features in each iteration. This results in a refined subset of features that enhances model accuracy [33, 161, 209, 211]. RFE is particularly effective for models with a large number of spectral bands. It is performed using a Python library *sklearn.feature_selection.RFE*[5]. RFE is used with any machine learning model as the estimator. Starting with all spectral bands, the model recursively eliminated the least essential features (lowest weight coefficients). However, it is computationally expensive for hyperspectral datasets because it repeatedly trains the model.

3. ***Pearson Correlation Coefficient (PCC):*** It measures the linear correlation between spectral bands and the target variables [186, 203–205]. Features with higher correlation are retained because these features contribute more to the model's predictive accuracy. Spectral analysis commonly uses PCC to identify bands by computing the correlation between each spectral band and the target variable. Bands with absolute correlation above a threshold (typically $r > 0.4$) are selected. This method is helpful for quick and interpretable feature selection. The main limitation is that PCC only captures linear relationships and is sensitive to noise, missing nonlinear spectral–soil interactions.

4. ***Competitive Adaptive Reweighted Sampling (CARS):*** A feature selection algorithm that iteratively selects features that are most informative about the target variable [153, 185, 207, 213]. It is particularly effective for spectral data where redundant wavelengths can reduce model accuracy. Less informative or redundant wavelengths are removed at each iteration, while those with stronger

[4] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

predictive power are retained. CARS is especially suitable for hyperspectral data, which typically contains hundreds of correlated and noisy wavelengths. However, CARS can be unstable and sensitive to noise, sometimes selecting different sets of features across runs.

5. ***Discriminant Analysis:*** In this technique, a linear discriminant function is used to predict the group membership of an observation [214, 215]. The linear discriminant process combines features, and the linear combination weights are determined by maximizing the between-group variance and minimizing the within-group conflict [93]. Although discriminant analysis is computationally efficient and interpretable, it assumes linear boundaries and equal covariance among classes, which may limit its applicability in complex, nonlinear spectral datasets. The assumption of linear separability limits performance when spectral properties exhibit nonlinear behaviour.

6. ***Least Absolute Shrinkage and Selection Operator (LASSO):*** This is a regression-based feature selection method that applies L1 regularization, effectively reducing the number of selected features by forcing some coefficients to zero [187, 205]. It is beneficial in eliminating redundant or less informative bands, leading to a more efficient and interpretable model. LASSO is applied using a Python library *sklearn.linear_model.Lasso*[6]. The LassoCV model is chosen to determine the optimal alpha value through a cross-validation process to ensure the best regularization strength. The training model keeps features with non-zero coefficients. This method is effective in handling high-dimensional spectral data. But LASSO can become unstable when predictors are highly correlated, often selecting only one from a correlated group.

7. ***Ant Colony Optimization (ACO):*** It is applied to combinatorial optimization issues to determine the optimal or nearly optimal solutions [149, 216]. By creating paths through a graph representing the features, ACO iteratively builds solutions for feature selection, biasing the search towards promising subsets based on pheromone trails [149]. Its ability to handle high-dimensional search spaces

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html`

and incorporate both global and local search makes it well-suited for selecting informative wavelengths from complex hyperspectral data. However, ACO is computationally intensive and may converge slowly for large spectral datasets.

8. ***Successive Projections Algorithm (SPA):*** This is a forward selection algorithm that iteratively selects features to minimize redundancy and maximize information content [185, 203, 212, 213]. It is particularly effective for hyperspectral data with highly correlated bands. SPA ensures that the selected features are minimally dependent on each other, which improves model robustness. It selects minimally collinear bands by maximizing vector projections. Its limitation is that SPA focuses on reducing multicollinearity but does not explicitly optimize for predictive relevance.

9. ***XGBoost Feature Importance:*** It is a robust gradient-boosting algorithm that provides built-in feature importance based on its decision trees [188]. By analyzing the gain, coverage, or frequency of features used in the trees, XGBoost can rank features according to the contribution to model accuracy. This method is particularly effective for large datasets and provides valuable insights into feature relevance [217, 218]. XGBoost models are trained separately for each target property. The 'gain' metric extracts feature importance scores. Top brands are selected based on the importance ranking. This approach helps uncover nonlinear dependencies and compares them against linear techniques, such as PCC and LASSO. However, XGBoost importance can be biased toward high-cardinality or high-variance features, and different metrics (gain/cover/weight) may yield inconsistent rankings.

3.4 Classification of ML and DL Models

ML is a branch of AI that focuses on creating algorithms and statistical models that enable computers to learn from data, improving performance at a given task without being explicitly programmed. Table 3.10 shows an overview of previous studies employing spectral data to predict soil nutrients using ML models. It presents details such as the type of spectral sensors and ML techniques adopted in different studies.

This comparison helps to understand the progress, challenges, and effectiveness of various machine learning approaches in predicting soil nutrients.

Regression techniques are chosen for soil nutrient prediction using spectral data since these techniques can accurately characterize the complex, nonlinear relationship between spectral data and soil nutrient contents [199]. It handles multidimensional data, provides quantitative predictions essential for precision farming, and offers interpretability for subject matter experts. Additionally, it can be systematically verified and enhanced, making it a trustworthy instrument for precise soil nutrient prediction and management [219]. ML/DL regression models are statistical methods used in data analysis to comprehend and forecast the relationship between a single or a set of independent variables (also called predictors or features) and a dependent variable (the goal or outcome) [220]. These models quantify and analyze the relationship between the spectral data (independent variables) and soil nutrient levels (dependent variables) [150]. The goal is to develop a model that predicts nutrient concentrations based on spectral information.

3.4.1 List of ML Models

Machine learning models provide powerful tools for analyzing soil nutrients. These models utilize computational learning to identify relationships that improve prediction performance across various soil types. The following are some typical ML techniques for analyzing soil nutrients [286]:-

1. ***Random Forest (RF):*** It employs ensemble learning by constructing multiple decision trees during training and averaging their predictions for the final output [161, 248, 249, 258]. RF handles high-dimensional data well, is resistant to overfitting, and automatically provides feature importance, which is useful for selecting spectral bands. Its robustness to noise and capability to model nonlinear relationships make it a strong choice for soil nutrient prediction tasks involving complex spectral data.

2. ***Support Vector Machine (SVM):*** It works by finding the optimal hyperplane that separates different classes with maximum margin; in regression (SVR), and fits a function within a defined error margin while minimizing model complexity

Table 3.10: Previous studies on ML-based soil nutrients prediction

Previous Work	HS/MS	Region	Nutrients	Spectral Data Tool	Model	Results
Chlouveraki et al. (2025) [86]	HS	Greece	pH, EC, SOM, CaCO3, N, P, K, Ca, Mg, Fe, Cu, Mn, Zn, B	Ground Based: Spectrometer, Private soil data	PCR, PLSR, MLP	R^2 ranges from 0.57 to 0.73 with MLP
Xu et al. (2024) [87]	MS	China	Soil pH and P	Ground Based: Spectrometer, Private soil data	Cubist, SVM, XGBoost	R^2 0.97 and 0.87 for soil pH and P with SVM
Wang et al. (2024) [88]	MS	China	Soil pH, SOM, N, P, K	Ground Based: Spectrometer, Private soil data	PLSR, Cubist, RF, SVM, CNN	R^2>0.39 with RF
El Harti et al. (2024) [89]	MS	Spain	SOC	UAV, Private soil data	ANN, SVM, RF, MLR	R^2=0.97 with RF
Pellikka et al. (2023) [221]	HS	Kenya	SOC and N	Ground Based: Spectrometer, Private soil data	LASSO Regression, GPR, RF, PLSR, CNN	R^2 0.78, 0.70 for SOC and N with LASSO
Singha et al. (2023) [222]	HS	India	SOC, pH, EC, N, P, K	Ground Based: Spectrometer, Private soil data	PLSR, SVR	Accuracy 87.2% and 88.9% with PLSR and SVR
Zeraatpisheh et al. (2023) [223]	MS	USA	SOC	Spaceborne: Landsat-8, Sentinel-2, Private soil data	RF, SVM, KNN, Cubist, Gradient Boosting	R^2 0.54 with RF
Abdoli et al. (2023) [66]	MS	Iran	SOC	Spaceborne: Landsat-8, Sentinel-2, MODIS, Private soil data	ANN, SVR, Genetic Expression Programming	R^2 0.627 with SVR
Monsalve et al. (2023) [224]	HS	Colombia	SOC and N	Airborne: HySpex, Private soil data	RF, SVM	R^2 0.92 (SOC), 0.87 (N) with SVM
Yan et al. (2023) [67]	HS	China	SOM	UAV, Private soil data	RF, Ordinary Kriging	R^2 0.57 with RF
Angelopoulou et al. (2023) [225]	HS	Greece	SOM and Ca	Airborne: HySpex, Spaceborne: PRISMA, Private soil data	PLSR, Cubist, RF, SVM	R^2 0.79 (SOM), 0.82 (Ca) with Cubist
Wu et al. (2023) [226]	MS	China	SOC	Spaceborne: Landsat-8, Private soil data	GBDT, Linear Regression	R^2 0.147 with GBDT
Luo et al. (2023) [227]	MS	China	SOM	Spaceborne: Sentinel-2, Private soil data	RF	R^2 0.445
Zhang et al. (2023) [228]	MS	China	SOC	Spaceborne: Sentinel-1, Sentinel-2, Private soil data	XGBRF	R^2 0.66
Chang et al. (2023) [229]	HS	China	SOC	UAV: GaiaSorter-dual camera, Private soil data	PLSR	R^2 0.61
Liu et al. (2023) [230]	MS	Pakistan	N, P, K, pH, SOM	Spaceborne: Landsat-8, MODIS, Private soil data	RF, MLR	R^2 0.94
Wang et al. (2022) [231]	MS	China	Soil pH	Ground Based: Spectrometer, Private soil data	PLSR, SVM, ELM	R^2 0.86 with ELM
Wang et al. (2022) [232]	Both	China	SOC	Airborne and Spaceborne sensors, Private soil data	PLSR, RF, Ridge, KNN, ANN, CNN, LSTM	R^2 0.96 with LSTM

Note: HS-Hyperspectral, MS-Multispectral

Table 3.10: Previous studies on ML-based soil nutrients prediction

Previous Work	HS/MS	Region	Nutrients	Spectral Data Tool	Model	Results
Garosi et al. (2022) [233]	MS	Iran	SOC	Ground Based: Spectrometer, Private soil data	RF, SVM	R^2 0.32 with SVM
Lu et al. (2022) [234]	MS	China	SOM	Ground Based: Spectrometer, Private soil data	Bayesian Maximum Entropy-RF	R^2 0.89
Heil et al. (2022) [235]	MS	Germany	SOM	UAV, Private soil data	RF	RMSE 0.13
Luo et al. (2022) [236]	MS	China	SOM	Spaceborne: Landsat-8, Private soil data	RF	R^2 0.705
Zeraatpisheh et al. (2022) [155]	MS	China	SOC	Ground Based: Spectrometer, Private soil data	RF, Cubist, PLSR, SVM	R^2 0.35 with Cubist
Shafizadeh et al. (2022) [237]	MS	Iran	SOC	Spaceborne: Sentinel-1, Sentinel-2 [238], Private soil data	RF, SVR	R^2 57.59% with SVR
Rodriguez et al. (2022) [239]	MS	Spain	N, SOM	Ground Based: Spectrometer, Private soil data	PLSR, RF	R^2 0.69 (N), 0.73 (SOM) with PLSR
Clingensmith et al. (2022) [240]	HS	USA	SOC, N, S, Ca, pH	Ground Based: Spectrometer, Private soil data	Cubist and PLSR	$R^2 > 0.9$ (Cubist)
Xie et al. (2022) [241]	MS	China	SOC	Spaceborne: Sentinel-2, Landsat-8, Private soil data	RF, XGBoost	$R^2 = 0.84$ (XGBoost)
Bulan et al. (2022) [242]	HS	Indonesia	N, P, K	Ground Based: Spectrometer, Private soil data	SVR, PLS-ANN, GBRT	RPD = 2.64, 3.93, 2.38 (GBRT)
Kuzu et al. (2022) [195]	HS	Poland	K, Mg, P, pH	Spaceborne: Intuition-1 (HYPERVIEW)	Hybrid (RF+KNN)	MSE = 0.783
Wang et al. (2022) [243]	HS	China	OM, P, K	Ground Based: Spectrometer, Private soil data	Ba-AdaBoost	$R^2 = 0.98, 0.96, 0.90$
Levi et al. (2022) [244]	HS	Israel	P, K, Ca, Mg, Cl, SOM	Airborne: AisaFenix sensor, Private soil data	SVR	$R^2 = 0.87$
Wei et al. (2022) [132]	HS	Brazil	P, K, Ca, Mn, pH, SOM	Ground Based: Spectrometer, Private soil data	MLR	$R^2 = 0.33$–0.96
Riad et al. (2022) [200]	HS	Bangladesh	P, K, B, Ca, Mg, Mn	Spaceborne: Glovis Earth Explorer, Private soil data	Hybrid	F1 = 80–88%
Gasmi et al. (2022) [245]	HS	Morocco	P, K, SOM	Spaceborne: Prisma, Private soil data	RF	$R^2 = 0.44$–0.69
Zhou et al. (2021) [246]	MS	China	N	Spaceborne: Landsat-8, Private soil data	PLSR, BPNN, SVM	$R^2 = 0.676$ (SVM)
Meng et al. (2021) [161]	HS	China	SOM	Spaceborne: Gaofen-5, Private soil data	RF	$R^2 = 0.84$
Yang et al. (2021) [29]	HS	China	N, SOM	UAV, Private soil data	PSO-ELM	$R^2 = 0.73$ (SOM)
Sankey et al. (2021) [247]	Both	Mexico	N, SOC	UAV, Private soil data	CART	Accuracy = 87%
Pechanec et al. (2021) [206]	HS	Czech Republic	N	Ground Based: Spectrometer, Private soil data	PLSR	$R^2 = 0.44$

Note: HS - Hyperspectral, MS - Multispectral

Table 3.10: Previous studies on ML-based soil nutrients prediction

Previous Work	HS/MS	Region	Nutrients	Spectral Data Tool	Model	Results
Yu et al. (2021) [248]	MS	Canada	N	UAV, Private soil data	RF, SVR	R^2 = 0.73
Wang et al. (2021) [249]	MS	China	N, P, K, SOM	Ground Based: Spectrometer, Private soil data	RF	Accuracy = 95%
Bao et al. (2021a) [250]	HS	China	SOM, N	UAV, Private soil data	ELM	R^2 = 0.73, 0.63
Xu et al. (2021) [251]	HS	China	N	Ground Based: Spectrometer, Private soil data	ANN, Cubist, KNN, SVM, XGBoost, PLSR	R^2 = 0.94 (SVM)
Hossen et al. (2021) [252]	MS	USA	N	UAV, Private soil data	MLP, SVM	SVR least error (44.46)
Bao et al. (2021b) [253]	MS	China	SOM	Ground Based: Spectrometer, Private soil data	SVM, PLSR, RF	RMSE = 4.34 (RF)
Zhang et al. (2021) [254]	MS	China	SOM	Spaceborne: Sentinel-2, MODIS, Private soil data	RF, ANN, SVR	R^2 = 0.67 (RF)
Zhou et al. (2021a) [255]	MS	China	N	Ground Based: Spectrometer, Private soil data	SVM	R^2 = 0.89
Tsakiridis et al. (2021) [256]	HS	Europe	SOC	Ground Based: Spectrometer, Private soil data	PLSR, SVR	R^2 = 0.77 (SVR)
Wang et al. (2021a) [257]	MS	China	SOC	Spaceborne: Sentinel-1,2, Private soil data	XGBoost	R^2 = 0.86
Zhou et al. (2021b) [246]	MS	Switzerland	SOC	Spaceborne: Sentinel-2,3, Landsat-8, Private soil data	SVM, BRT, RF	R^2 ↑ 29.1% (BRT)
He et al. (2021) [70]	MS	China	SOC	Spaceborne: Sentinel-2, Private soil data	RF	R^2 ↑ 171%
Shi et al. (2021) [258]	HS	Morocco	SOC	Spaceborne: Prisma, Private soil data	RF	R^2 = 0.92
Wang et al. (2021b) [259]	MS	China	SOM	Spaceborne: Sentinel-2, Private soil data	PLSR, SVR, ANN	R^2 = 0.723 (ANN)
Xu et al. (2021) [260]	HS	China	SOC	Ground Based: Spectrometer, Private soil data	SVM, PLSR, WNN	R^2 = 0.92 (SVM)
Ghazali et al. (2020) [32]	MS	Indonesia	pH	Spaceborne: Landsat-8, Private soil data	MLR	R^2 = 0.81
Lin et al. (2020) [261]	HS	China	N	Ground Based: Spectrometer, Private soil data	SCLM, PLSR	R^2 = 0.88 (SCLM)
Zhang et al. (2020) [219]	HS	China	SOM	Ground Based: Spectrometer, Private soil data	PLS-SVM	R^2 = 0.846
Hong et al. (2020) [262]	HS	USA	SOC	Airborne: Headwall Micro-Hyperspec, Private soil data	RF	R^2 = 0.66
Zhou et al. (2020a) [263]	MS	Europe	SOC, N	Spaceborne: Sentinel-1,2, Private soil data	RF, BRT, SVM, CART	R^2 ↑ 12.8–18.8%
Zhou et al. (2020b) [264]	MS	China	SOC	Spaceborne: Sentinel-1, Private soil data	RF, BRT	R^2 = 0.75 with RF

Note: HS - Hyperspectral, MS - Multispectral

Table 3.10: Previous studies on ML-based soil nutrients prediction

Previous Work	HS/MS	Region	Nutrients	Spectral Data Tool	Model	Results
Wang et al. (2020) [265]	MS	Spain	SOC	Spaceborne: Landsat-8, Private soil data	RF	R^2 = 0.59
Yang et al. (2019) [35]	HS	China	SOM, pH	Ground Based: Spectrometer, Private soil data	PLSR, SVM, ELM, Cubist	ELM R^2 = 0.87 (SOM), 0.85 (pH)
Ballabio et al. (2019) [266]	MS	Europe	N, P, K, $CaCO_3$, CEC, pH	Spaceborne: MODIS, Private soil data	GPR	RMSE: N 1.97, P 17.52, K 121.89, $CaCO_3$ 78.29, pH 0.78, CEC 11.02
Castaldi et al. (2019) [201]	MS	Europe	SOC	Spaceborne: Sentinel-2, Private soil data	Exp. Regression Model	RPD = 2.9
Hong et al. (2019) [267]	HS	China	Lead, Zinc	Ground Based: Spectrometer, Private soil data	PLSR, RF	R^2 = 0.82 (RF)
Dou et al. (2019) [268]	MS	China	SOM	Spaceborne: MODIS, Private soil data	Stepwise MLR	R^2 = 0.725
Li et al. (2019) [269]	HS	China	N	Ground-Based: Spectrometer, Private soil data	PLS, ELM	R^2 = 0.94 (ELM)
Moura et al. (2019) [270]	MS	Brazil	SOC	Ground Based: Spectrometer, Private soil data	MLR, PLSR, SVM, RF	R^2 = 0.74 (PLSR)
Dangal et al. (2019) [52]	HS	USA	Ca, Fe, SOC, pH	Ground Based: Spectrometer, Private soil data	PLSR	R^2 = 0.62
Ballabio et al. (2018) [51]	MS	Europe	Cu	Spaceborne: MODIS, Private soil data	GPR	66% explained deviance
Castaldi et al. (2018) [271]	MS	Belgium, Luxembourg	SOC	Ground Based: Spectrometer, Private soil data	PLSR	RMSE = 3.7, RPD = 2.54
Song et al. (2018) [272]	HS	China	N, P, K	Spaceborne: Environment 1A Satellite, Private soil data	SLR, SVM, RF, BPNN, BPNNOK	RMSE: N 0.292, P 29.62, K 49.67
Hong et al. (2018) [273]	HS	China	SOM	Ground Based: Spectrometer, Private soil data	SVM, ELM	R^2 = 0.83 (ELM)
Yue et al. (2018) [274]	HS	China	SOM	Ground Based: Spectrometer, Private soil data	RF	R^2 = 0.838
Xu et al. (2018a) [275]	MS	China	SOM, N, P, K	Ground Based: Spectrometer, Private soil data	PLSR, BPNN, SVM	R^2: SOM 0.88, N 0.86, P 0.76, K 0.65
Xu et al. (2018b) [276]	MS	China	pH, N, P, K, SOM, Fe	Ground Based: Spectrometer, Private soil data	PLSR	R^2 = 0.70–0.90
Ng et al. (2018) [277]	HS	Europe, Australia	SOC, pH	Ground Based: Spectrometer, Private soil data	PLSR, Cubist, K-means	$R^2 > 0.70$ (K-means)
Qi et al. (2018) [278]	HS	China	N, P, K	Ground Based: Spectrometer, Private soil data	PLSR	RPD: P 2.23, K 1.47, N low
Milovský et al. (2017) [279]	MS	Croatia	SOC	Ground Based: Spectrometer, Private soil data	PLSR	R^2 = 0.83

Note: HS - Hyperspectral, MS - Multispectral

Table 3.10: Previous studies on ML-based soil nutrients prediction

Previous Work	HS/MS	Region	Nutrients	Spectral Data Tool	Model	Results
Jia et al. (2017) [280]	HS	China	N	Ground Based: Spectrometer, Private soil data	PLSR, SVM	Accuracy = 91.8%
Zhang et al. (2016) [281]	MS	China	N	Ground Based: Spectrometer, Private soil data	PLSR, SVM	R^2 = 0.823 (SVM)
Morellos et al. (2016) [282]	MS	Germany	SOC, N	Ground Based: Spectrometer, Private soil data	PLSR, SVM, Cubist	RPD: SOC 2.20 (SVM), N 1.96 (Cubist)
Srivastava et al. (2015) [283]	HS	India	SOC	Ground Based: Spectrometer, Private soil data	PLSR	R^2 = 0.81
Vohland et al. (2014) [284]	HS	Germany	pH, SOC, N	Ground Based: Spectrometer, Private soil data	PLSR, CARS-PLSR	CARS-PLSR RPD $>$ 3
Nocita et al. (2014) [285]	MS	Europe	SOC	Ground Based: Spectrometer, Private soil data	PLSR	RMSE = 3.6

Note: HS - Hyperspectral, MS - Multispectral

[35, 242, 259]. SVM effectively handles high-dimensional data and can model non-linear patterns using kernel functions. However, it is sensitive to the choice of kernel and parameters, and its computational cost can be high for large datasets. Despite this, SVM remains a reliable tool for accurately estimating soil parameters using spectral features.

3. ***Extreme Gradient Boost (XGBoost):*** It is an advanced boosting technique that builds decision trees sequentially, where each new tree corrects the residuals of the previous ones [241, 257]. XGBoost includes regularization to prevent overfitting and efficiently handles missing values and sparse data. Its scalability and high prediction accuracy are well-suited for hyperspectral datasets with common noise and high dimensionality. Additionally, XGBoost provides feature importance scores that aid in selecting the optimal wavelength.

4. ***Multiple Linear Regression (MLR):*** It models the linear relationship between a dependent variable and multiple independent variables. MLR is simple, interpretable, and computationally efficient, but assumes linearity, homoscedasticity, and no multicollinearity—conditions often violated in complex spectral data [32, 132, 230, 270]. Despite its limitations, MLR can serve as a baseline model for predicting soil nutrients and is particularly helpful when the relationships between spectral bands and nutrients are approximately linear.

5. ***Partial Least Square Regression (PLSR):*** It reduces the predictors to a smaller set of uncorrelated components and performs regression on them, making it highly suitable for high-dimensional, multicollinear data like hyperspectral reflectance [35,242,259]. PLSR captures the covariance structure between the predictors and the response, thereby enhancing the model's robustness and predictive accuracy. It is widely used in chemometrics and remote sensing, particularly when the number of wavelengths exceeds the number of samples.

6. ***Cubist:*** A rule-based regression model integrates decision tree logic with linear regression models at the terminal nodes. It creates piecewise models that can adapt to local patterns in the data and refine predictions through instance-based corrections [87, 88, 223]. This hybrid structure enables Cubist to capture both global trends and local variations in complex datasets, such as soil spectra, resulting in high prediction accuracy for soil nutrients.

7. ***Gaussian Process Regression (GPR):*** A non-parametric Bayesian regression technique provides predictions and uncertainty predictions by defining a distribution over functions [221]. It is beneficial when the dataset is small, but the uncertainty quantification is critical. However, its computational complexity grows cubically with the number of samples, which limits its scalability. GPR is valuable for predicting soil nutrients when prediction confidence is as important as accuracy.

8. ***Extreme Learning Machine (ELM):*** It is a type of single-layer feedforward neural network where the input weights are randomly initialized and only the output weights are learned, resulting in speedy training times [231, 250, 269]. ELM is suitable for large datasets and applications requiring rapid model development. Despite its simplicity, ELM can approximate complex nonlinear relationships between spectral bands and soil nutrients, making it a valuable tool in high-throughput soil analysis scenarios.

3.4.2 List of DL Models

DL is a branch of ML that concentrates on teaching artificial neural networks, deep

Table 3.11: Previous studies on DL-based soil nutrients prediction

Previous Work	HS/MS	Region	Nutrients	Spectral Data Tool	Model	Results
Qi et al. (2025) [287]	HS	China	SOM, N	Ground Based: Spectrometer, Private soil data	BPNN	R^2 = 0.980 (SOM), 0.972 (N)
Bolanos et al. (2024) [18]	HS	Colombia	P	Ground Based: Spectrometer, Univalle [6]	ANN	R^2 = 0.867
Li et al. (2024) [288]	HS	China	SOC	UAV: GaiaField Pro-V10 HSI, Private soil data	LSTM	RPD = 47.36%
Gacek et al. (2024) [192]	HS	Poland	P, K, Mg, pH	Spaceborne: Intuition-1 satellite, HYPERVIEW [4]	SVR, RF, CNN2D	CNN2D Score: P 0.92, K 0.98, Mg 0.88; SVR/RF Score for pH 0.93
Samarinas et al. (2023) [289]	MS	Lithuania	SOC	Spaceborne: Sentinel-2, Private soil data	CNN	R^2 = 0.51
Gao et al. (2023) [290]	MS	China	SOC	Spaceborne: Sentinel-2, Private soil data	BPNN	R^2 = 0.773
Piccoli et al. (2023) [199]	HS	Europe	N, P, K, soil pH, SOC	Spaceborne: PRISMA, Private soil data	CNN	Avg RMSE = 0.1015
Zhao et al. (2023) [189]	HS	Africa and Europe	N, P, K, SOC, soil pH, Ca	Ground Based: Spectrometer, LUCAS-2009 [291], AfSIS [190]	CNN	Best R^2 (soil pH) = 0.94
Liu et al. (2022) [292]	HS	USA	SOC	Ground Based: Spectrometer, Private soil data	BPNN, DNN	R^2 = 0.81 (DNN)
Zelenka et al. (2022) [17]	HS	Poland	P, K, Mg, pH	Spaceborne: Intuition-1 satellite, HYPERVIEW [4]	SpectralNet3D	Scores: P 0.976, K 0.439, Mg 0.814, pH 0.889
Tripathi et al. (2022) [5]	MS	India	Soil pH, SOC	Spaceborne: Sentinel-2, Private soil data	MLP	R^2 = 0.684
Wang et al. (2022) [293]	HS	China	Cu	Ground Based: Spectrometer, Private soil data	TabNet	R^2 = 0.942
Devidhanshrii et al. (2022) [294]	HS	India	N, P, K, soil pH, SOM	Ground Based: Spectrometer, Private soil data	CNN	Accuracy = 99%
Zhao et al. (2022) [190]	HS	Africa	SOC	Ground Based: Spectrometer, LUCAS-2009 [291], AfSIS	CNN	R^2 = 0.93
Pham et al. (2021) [295]	MS	58 countries	SOC	Ground Based: Spectrometer, Private soil data	CNN-based RDNet, DenseNet	R^2 = 0.86
Zhao et al. (2021) [65]	HS	Africa and Europe	SOC	Ground Based: Spectrometer, LUCAS-2009 [291], AfSIS	DNN, LSTM, CNN	Best R^2 = 0.93
Tsimpouris et al. (2021) [296]	MS	Europe	Soil pH, $CaCO_3$, N	Ground Based: Spectrometer, Private soil data	Autoencoders	RMSE decreased by 9.9%
Zhong et al. (2021) [191]	HS	Europe	N, P, K, OC, soil pH, CEC, $CaCO_3$	Ground Based: Spectrometer, Private soil data	CNN	R^2 up to 0.96 (OC, pH, N, etc.)
Wang et al. (2021) [297]	HS	Europe	N	Ground Based: Spectrometer, Private soil data	CNN	R^2 = 0.93, RMSE = 0.97, RPD = 3.85
Patel et al. (2020) [298]	HS	India	N	Ground Based: Spectrometer, Private soil data	MLP	R^2 = 0.953

Note: HS - Hyperspectral, MS - Multispectral

Table 3.11: Previous studies on DL-based soil nutrients prediction

Previous Work	HS/MS	Region	Nutrients	Spectral Data Tool	Model	Results
Reda et al. (2020) [299]	MS	Morocco	P	Ground Based: Spectrometer, Private soil data	BPNN	R^2 = 0.86
Zhang et al. (2020) [300]	MS	China	P	Spaceborne: Landsat-8, Private soil data	ANN	R^2 = 0.63
Tsakiridis et al. (2020) [301]	HS	Europe	SOC, N	Ground Based: Spectrometer, LUCAS-2009 [291]	CNN	SOC R^2 = 0.86, N R^2 = 0.83
Yang et al. (2020) [302]	MS	Europe	SOC, N, soil pH	Ground Based: Spectrometer, Private soil data	CNN, RNN, CCNVR	CCNVR R^2: SOC 0.73, N 0.70, soil pH 0.86
Padarian et al. (2019) [303]	HS	Europe	SOC, soil pH, N	Ground Based: Spectrometer, Private soil data	CNN, PLS, Cubist	CNN reduced error by 87% (vs. PLS) and 62% (vs. Cubist)
Fernandes et al. (2019) [304]	MS	Brazil	SOM	Ground Based: Spectrometer, Private soil data	ANN	R^2 = 0.76
Ng et al. (2020) [305]	HS	Brazil	SOC	Ground Based: Spectrometer, Private soil data	CNN	R^2 = 0.69
Ng et al. (2019) [306]	HS	USA	SOC, soil pH	Ground Based: Spectrometer, Private soil data	CNN	R^2 = 0.95–0.98
Tsakiridis et al. (2019a) [307]	MS	Europe	SOM	Ground Based: Spectrometer, Private soil data	Genetic Algorithm	R^2 = 0.76
Tsakiridis et al. (2019b) [308]	HS	Europe	SOC	Ground Based: Spectrometer, Private soil data	Evolutionary Fuzzy Rule-based System	R^2 = 0.77
Tsakiridis et al. (2019c) [309]	MS	Europe	SOC	Ground Based: Spectrometer, Private soil data	Evolutionary Fuzzy Rule-based System	R^2 = 0.91
Singh et al. (2019) [310]	HS	Europe	SOC, N, soil pH	Ground Based: Spectrometer, Private soil data	LSTM	R^2 = 0.94
Wang et al. (2018) [311]	Both	China	SOM	Ground Based: Spectrometer and Landsat-8, Private soil data	BPNN	RPD = 2.26
Sila et al. (2016) [312]	MS	Africa	Soil pH, SOC	Spaceborne: Sentinel-2, Private soil data	SOM	RMSE = 1.25

Note: HS - Hyperspectral, MS - Multispectral

neural networks, to process and learn from massive volumes of data. Table 3.11 presents an overview of earlier studies that employed spectral data to predict soil nutrients using DL models. It summarizes the datasets, deep learning architectures, and target nutrients used, highlighting the strengths and limitations of each approach. This thesis clearly compares how different DL models have been applied for soil nutrient prediction and the reported performance. Some DL models for analyzing soil nutrients include ([47, 131]):-

1. ***Artificial Neural Network (ANN):*** It is a computational model inspired by the

structure of the human brain, consisting of interconnected nodes (neurons) organized in layers [5, 298]. ANN learns patterns and complex relationships in input data through training and backpropagation. It is versatile and can be used for classification and regression tasks, including predicting soil nutrients based on spectral inputs. MLP, a common form of ANN, has shown promising results in modeling the nonlinear relationships between hyperspectral features and soil nutrients.

2. ***Convolutional Neural Network (CNN):*** It is a deep learning model specialized for extracting spatial features from grid-like structured data. In hyperspectral soil analysis, CNNs are applied to learn spatial-spectral correlations by using convolutional filters that detect patterns such as edges, gradients, or spectral shifts [191, 199, 294]. This model is particularly effective in identifying high-level features relevant to soil nutrients across a wide range of wavelengths. Variants like DenseNet and RDNet have been successfully employed to predict SOC and other nutrients.

3. ***Backpropagation Neural Network (BPNN):*** It is a supervised learning model where the error between predicted and actual outputs is propagated backward through the network to update the weights [290]. BPNN is particularly suited for regression tasks, such as predicting continuous values like soil nutrient concentrations. By learning complex input-output mappings, it can model the nonlinear dependencies between hyperspectral reflectance and soil chemical properties.

4. ***TabNet:*** It is a deep learning model tailored for tabular data, combining neural network structures with sequential attention mechanisms [293]. TabNet learns feature selection dynamically at each decision step, which helps understand which features (e.g., spectral bands) are most relevant for prediction. Its interpretability, ability to handle high-dimensional data, and end-to-end learning capability make it well-suited for estimating soil nutrients using hyperspectral data.

5. ***Deep Neural Network (DNN):*** It refers to neural networks with multiple hidden layers, enabling the model to learn hierarchical and abstract data representations

[292, 313]. DNN captures complex nonlinear relationships between inputs and outputs, making it effective for modeling the intricate patterns found in spectral soil data. It is known for its strong generalization ability when trained on sufficiently large datasets and has shown reliable performance in estimating SOC and other soil nutrients.

6. ***Long Short-Term Memory (LSTM):*** It is a type of Recurrent neural network (RNN) that can retain information over long steps using gated memory cells, making it capable of modeling sequential dependencies [288, 310]. In hyperspectral analysis, where data can be treated as a sequence of spectral bands, LSTM helps capture the relationships across wavelengths more effectively than traditional feedforward networks. This ability to model temporal/spectral dependencies makes LSTM a powerful tool for predicting soil nutrients.

7. ***Self-Attention Mechanism:***: It captures relationships among different wavelength bands in hyperspectral data by assigning weights according to their relevance. In soil nutrient prediction, it helps identify informative spectral regions and learn long-range dependencies, enabling the model to extract meaningful spectral features for accurate prediction [314, 315]. This mechanism enables the model to learn complex spectral interactions that may be associated with variations in soil nutrient concentrations.

3.4.3 Contribution of ML Techniques in Soil Nutrient Prediction

The previous work shows that ML has been widely used for soil nutrient prediction using spectral data. Table 3.10 summarizes ML approaches used with multispectral and hyperspectral data for soil nutrient prediction. These methods focus on achieving high regression accuracies. This review focuses on the ML models used and the corresponding performance indicators. The literature presents numerous ML models adapted to meet the study's unique requirements.

Researchers frequently use RF, SVR, MLR, and PLSR to forecast soil nutrients. For example, RF and multiple linear regression achieved an impressive R^2 of 0.94 in predicting soil nutrient levels in a study by Liu et al. [230]. The gradient boosting

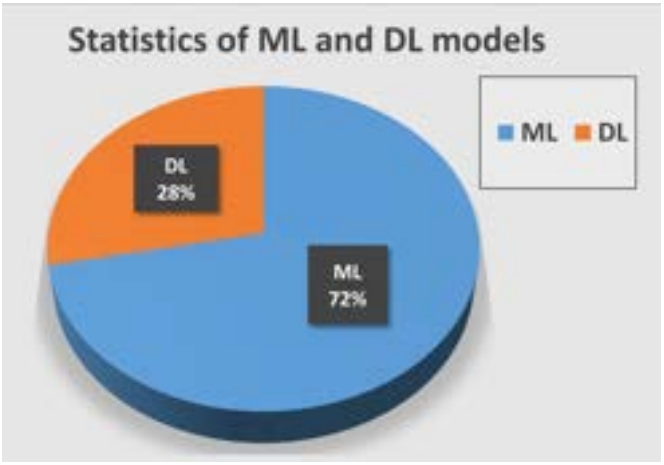

Figure 3.7: Usage of ML and DL models in previous studies for soil nutrient prediction

regressor used by Bulan et al. [242] achieved RPD values of 2.64, 3.93, and 2.38 for N, P, and K, respectively. Furthermore, Wang et al. [243], demonstrated R^2 values of 0.98, 0.96, and 0.90 for SOM, P, and K prediction, respectively, showing the effectiveness of hyperspectral data. In the meantime, PSO with Empirical Line Method (ELM) was used by Yang et al. [29] to obtain R^2 values of 0.73 and 0.63 for SOM and N. These findings demonstrate the vast potential of ML models for spectral data-based soil nutrient prediction. The versatility and robust performance metrics make them valuable tools for environmental monitoring and precision agriculture. The researchers anticipate that as sophisticated machine-learning techniques emerge, soil nutrient prediction algorithms will continue to be refined and developed.

3.4.4 Contribution of DL Techniques in Soil Nutrient Prediction

DL has advanced quickly in the last few years. This technique enables a computational model with several processing layers to learn different levels of abstraction in data representations [316]. More feature information may be obtained via DL, leading to better and more consistent results [317]. According to the literature, scientists have used multispectral and hyperspectral data to apply some DL models for soil nutrient prediction, as shown in Table 3.11. In this thesis, the researchers explore the dynamic landscape of DL techniques for soil nutrient prediction using spectral data from multiple geographies and various data-collection methodologies.

Previous research has demonstrated the possibilities of DL models, with a particular

emphasis on the efficacy. These studies utilized CNNs with notable success, primarily because CNNs exhibit exceptional performance in various categories of soil nutrients. Pham et al. [295] employed a CNN model and spectrometer to predict SOC levels across 58 countries, achieving an impressive R^2 value of 0.86. A multi-nutrient prediction for components such as N, P, K, SOC, soil pH, CEC, and $CaCO_3$ was also explored by Zhong et al. [191]. Using CNNs, these studies achieved great accuracy for different nutrients, with R^2 values ranging from 0.82 to 0.96. Moreover, DL methods have demonstrated effectiveness in diverse geographical locations, including China, Africa, and Europe. The ability to adapt to various environments highlights the adaptability and broad applicability. Patel et al. [298] predicted the nitrogen levels in Roorkee using Spectrometer data, and achieved an amazing R^2 value of 0.953 using an MLP model. Sila et al. [312] worked on the African region and got RMSE = 1.25 while concentrating on Sentinel-2 [238] data for soil pH and SOC prediction.

This thesis illuminates the potential of DL models in spectral data-based soil nutrient prediction. The strong performance and flexibility of DL methods across various geographies and nutrient categories will pave the way for the advancement of precision agriculture and environmental monitoring. As DL models continue to evolve, this work aims to drive ongoing innovation and improvement in soil nutrient prediction, promising bright prospects for future developments. Figure 3.7 shows that the previous studies mainly used traditional ML models for soil nutrient prediction, and there is limited coverage of DL models for soil nutrient prediction using spectral data. It indicates a potential opportunity to explore advanced DL techniques for improving prediction accuracy and capturing complex spectral–soil relationships.

3.5 Spectral Indices for Soil Nutrient Prediction

Spectral Indices are mathematical equations applied to each pixel image to highlight a specific phenomenon on the ground [318]. Some of the spectral indices that are used in agriculture are:

1. ***Vegetation Indices:*** These are mathematical combinations of spectral reflectance values, typically in the red and near-infrared regions, used to assess plant health, vigor, and biomass. Vegetation indices facilitate the monitoring of crop growth,

Table 3.12: Different types of Vegetation indices used for soil nutrient prediction.

Vegetation Indices	**Description**	**Formula**
NDVI (Normalized Difference Vegetation Index)	It measures the density of green vegetation in an area.	$NDVI = \frac{(\text{NIR}-\text{Red})}{(\text{NIR}+\text{Red})}$
EVI (Enhanced Vegetation Index)	A vegetation index that corrects for atmospheric influences and enhances sensitivity to canopy structure and health.	$\text{EVI} = Green \times \frac{\text{NIR}-\text{Red}}{\text{NIR}+C1\times\text{Red}-C2\times\text{Blue}+L}$
ARVI (Atmospherically Resistant Vegetation Index)	It corrects for atmospheric interference in NDVI.	$\text{ARVI} = \frac{\text{NIR}-(2X\text{RED}-\text{BLUE})}{\text{NIR}+(2X\text{RED}-\text{BLUE})}$
RDVI (Renormalized Difference Vegetation Index)	A variant of NDVI that aims to reduce soil background influences.	$\text{RDVI} = \frac{\text{NIR}-\text{Red}}{\sqrt{\text{NIR}+\text{Red}}}$
NPCRI (Normalized Pigment Chlorophyll Index)	Measures chlorophyll content in vegetation.	$\text{NPCRI} = \frac{\text{Red}-\text{Blue}}{\text{Red}+\text{Blue}}$
NDMI (Normalized Difference Moisture Index)	Emphasizes moisture content in vegetation and soil.	$\text{NDMI} = \frac{\text{NIR}-\text{SWIR}}{\text{NIR}+\text{SWIR}}$
SAVI (Soil Adjusted Vegetation Index)	It corrects for soil background influences in NDVI.	$\text{SAVI} = \left(\frac{\text{NIR}-\text{Red}}{\text{NIR}+\text{Red}+L}\right) \times (1+L)$
RVI (Ratio Vegetation Index)	A simple ratio of near-infrared to red reflectance.	$\text{RVI} = \frac{\text{NIR}}{\text{Red}}$
NDGI (Normalized Difference Greenness Index)	Related to the greenness or health of vegetation.	$\text{NDGI} = \frac{\text{NIR}-\text{Green}}{\text{NIR}+\text{Green}}$
NDRE (Normalized Difference Red-edge Index)	Measures variations in the red-edge region of the electromagnetic spectrum.	$\text{NDRE} = \frac{\text{NIR}-\text{RedEdge}}{\text{NIR}+\text{RedEdge}}$
GCI (Green Chlorophyll Index)	Measures chlorophyll content in vegetation.	$\text{GCI} = \frac{\text{NIR}}{\text{GREEN}} - 1$
NDWI (Normalized Difference Water Index)	It highlights the presence of water bodies and moisture content in vegetation.	$\text{NDWI} = \frac{\text{Green}-\text{NIR}}{\text{Green}+\text{NIR}}$
NDSI (Normalized Difference Snow Index)	It controls variance of two bands (one in the near-infrared or short-wave infrared and another one in the visible parts of the spectrum). This is useful for snow mapping.	$\text{NDSI} = \frac{\text{GREEN}-\text{SWIR1}}{\text{GREEN}+\text{SWIR1}}$
NDBI (Normalized Difference Built-up Index)	It detects built-up urban areas.	$\text{NDBI} = \frac{\text{SWIR}-\text{NIR}}{\text{SWIR}+\text{NIR}}$
HI (Hue Index)	It characterizes the dominant color or hue of an object in spectral imagery.	$\text{HI} = \frac{2\times Red-Green-Blue}{Green-Blue}$
RI (Redness Index)	Measure the relative amount of red light reflected by an object compared to other wavelengths.	$\text{RI} = \frac{\text{Red}\times\text{Red}}{\text{Green}\times\text{Green}}$

Abbreviations: NIR – Near-infrared; C1, C2 – Coefficients for the aerosol resistance term; L – Canopy background adjustment; SWIR – Short-wave infrared; SWIR1 – Short-wave infrared 1

detection of stress conditions, and yield prediction. Common examples include NDVI and EVI, which are widely used in precision agriculture and ecosystem

Table 3.13: Different types of soil and water indices used for soil nutrient prediction

Soil Indices	**Description**	**Formula**
Yellowness Index	It quantifies how yellow an object or material appears.	$\text{YI} = 0.723B3 - 0.597B4 + 0.206B6 - 0.278B9$
SCI (Soil Composition Index)	It likely relates to soil composition or properties.	$\text{SCI} = \frac{\text{SWIR1}-\text{NIR}}{\text{SWIR1}+\text{NIR}}$
Iron Index	It likely assesses the presence or concentration of iron minerals.	$\text{Iron Index} = \frac{\text{Red}}{\text{Blue}}$
Colour Index	CI is a measurement related to the color or pigmentation of a substance or object.	$\text{CI} = \frac{\text{Red}-\text{Green}}{\text{Red}+\text{Green}}$
Clay Index	It is likely used to assess soil properties such as clay content.	$\text{Clay Index} = \frac{\text{SWIR1}}{\text{SWIR}}$
BSI (Bare Soil Index)	It indicates the presence of bare soil or unvegetated surfaces.	$\text{BSI} = \frac{(\text{Red}+\text{SWIR})-(\text{NIR}+\text{Blue})}{(\text{Red}+\text{SWIR})+(\text{NIR}+\text{Blue})}$
Water Indices	**Description**	**Formula**
MSI (Moisture Stress Index)	Measures vegetation moisture stress.	$\text{MSI} = \frac{\text{MidIR}}{\text{NIR}}$

Table 3.14: Different types of temperature and thermal Indices used for soil nutrient prediction

Temperature and Thermal Indices	**Description**	**Formula**
LST (Land Surface Temperature)	It measures the temperature of the Earth's surface.	$\text{LST} = \frac{\text{BT}}{1+w\times\left(\frac{\text{BT}}{p}\right)\times\ln(E)}$
VTCI (Vegetation Temperature Cover Index)	Related to the temperature of vegetation cover.	$\text{VTCI} = \frac{\text{LST}_{\text{max NDVIi}}-\text{LST}(NDVIi)}{\text{LST}_{\text{max NDVIi}}-\text{LST}_{\text{min NDVIi}}}$
Brightness Index	It is likely related to the brightness or reflectance characteristics.	$\text{BI} = \sqrt{\frac{\frac{\text{Red}\times\text{Red}}{\text{Green}\times\text{Green}}}{2}}$
Wetness Index	WI determines the wetness or moisture content of the land surface.	$\text{WI} = \ln\left(\frac{a}{\tan b}\right)$

monitoring [319].

2. ***Soil Indices:*** These are designed to highlight specific soil nutrients such as moisture content, organic matter, salinity, or texture using spectral reflectance data. Soil indices enhance the sensitivity of spectral analysis to subtle soil characteristics, aiding in soil classification and fertility assessment [320].

3. ***Water Indices:*** It utilizes specific wavelengths to detect and monitor surface water bodies, predict turbidity, and evaluate water content in vegetation or soil. Water indices are crucial for mapping flood zones, drought monitoring, and wetland assessments, and are often derived from near-infrared and short-wave infrared bands [321].

4. ***Temperature and Thermal Indices:*** These are derived from thermal infrared data to monitor surface temperatures, heat stress, and urban heat islands. Temperature and thermal indices help to assess climate variability, support disaster response, and evaluate environmental and public health risks associated with extreme temperatures [322].

Tables 3.12, 3.13, and 3.14 present the categories of existing spectral indices used in previous studies for soil nutrient prediction [198, 313]. The formulas in Tables 3.12, 3.13, and 3.14 are defined according to the spectral characteristics of Landsat-8 [159] and Sentinel-2 [238]. These tables provide a comprehensive reference for selecting the appropriate indices based on the target soil nutrient and the applicability of different indices in predicting soil nutrients.

Table 3.15 shows the previous work by researchers who have implemented soil nutrient prediction using spectral indices, with the limitations. It summarizes the target nutrients, spectral indices, and the modeling approaches adopted in these studies. The table provides insights into the effectiveness and constraints of different spectral indices for accurate soil nutrient prediction. For instance, Gulhane et al. [30] employed spectral indices to predict parameters such as P, Fe, and soil pH in India, achieving impressive accuracy rates of 97%. However, a notable limitation is the need to explore the indices' correlation with other environmental factors, such as Normalized Difference Vegetation Index (NDVI), soil moisture, and temperature, which can influence soil nutrient predictions. Similarly, Babbar et al. [50] utilized multiple spectral indices to predict various soil parameters, including N, Mn, SOC, Ca, and Cu, in India. The results showcased high R^2 values within the range of 0.90 to 0.98. However, an inherent limitation is neglecting nonlinear relationships within the data, which can affect the accuracy of predictions. Although spectral indices have helped predict soil nutrients, still need to be more comprehensive and consider a broader range of environmental conditions. The study by Yuvaraj et al. [79] does not include temperature data, which can affect soil qualities. Hence, a broader range of spectral indices, including those related to soil moisture and temperature, must be incorporated to improve the accuracy and comprehensiveness of soil nutrient predictions. A few researchers also narrowed the understanding of multi-nutrient interactions in the soil by concentrating mainly on

Table 3.15: Previous studies on soil nutrient prediction through spectral indices

Previous Work	HS/MS	Region	Nutrients	Spectral Data Tool	Spectral Indices	Model and Results	Advantages	Limitations
Dhiman et al. (2023) [198]	MS	India	N, P, SOC, Ca, Mg	Spaceborne: Landsat-8, Sentinel-2, Private soil data	I4, I5, I6, I11, I13, I19	Model - GB; Results - R^2 N 0.98, P 0.84, SOC 0.97, Ca 0.98, Mg 0.95	Capturing the nuanced relationship between plant health and soil nutrients.	Temperature not considered.
Zayani et al. (2023) [313]	MS	France	SOC	Spaceborne: Sentinel-2, Private soil data	I4, I6, I10, I11, I22, I28	Model - DNN; Results - RPIQ 0.79	Explored the correlation of SOC with multiple spectral indices.	Temperature not considered; Only one nutrient targeted.
Gulhane et al. (2023) [30]	MS	India	P, Fe, soil pH	Spaceborne: Landsat-8, Sentinel-2, Private soil data	I1, I2, I3	Model - Iterative self organizing cluster unsupervised algorithm; Results - Accuracy 97%	Developed spectral indices for soil nutrients.	Correlation with other indices, such as NDVI, soil moisture, and temperature, has not been studied.
Babbar et al. (2022) [50]	MS	India	N, Mn, SOC, Ca, soil pH, Cu, Fe	Spaceborne: Landsat-8, Sentinel-2, Private soil data	I4, I5, I6, I7, I8, I9, I10, I11	Model - Multilinear Regression Analysis; Results - R^2 within 0.90–0.98	Explored soil nutrient and spectral indices correlations.	Ignores nonlinear relationships.
Yuvaraj et al. (2022) [79]	MS	Florida	SOM, soil pH	Spaceborne: Landsat-8, Sentinel-2, Private soil data	I4, I6, I12, I13, I14, I15, I16	Model - Linear Regression; Results - R^2 0.623	1. It explored SOM, soil pH, and spectral indices correlations.	1. Temperature not considered.
Wang et al. (2022) [323]	MS	China	SOM	Spaceborne: Sentinel-2, Private soil data	I10, I28	Model - RF; Results - R^2 0.59	1. Improves the accuracy of the SOM prediction while considering bare soil.	1. Exploration of limited spectral indices. 2. Only one nutrient targeted.
Yu et al. (2022) [324]	MS	Tibet	SOM	Ground Based: Spectrometer, Private soil data	I4, I20	Model - BPNN; Results - R^2 0.97	1. Explored correlation of LST, NDVI with SOM.	1. Exploration of limited spectral indices.
Rodriguez et al. (2022) [239]	MS	China	SOM	Ground Based: Spectrometer, Private soil data	I4, I5, I6, I11, I15, I29, I30	Model - RF; Results - R^2 0.57	1. Captured the nuanced relationship between plant health and SOM.	1. Temperature not considered. 2. Only one nutrient targeted.
Yu et al. (2021) [325]	MS	China	SOM	Spaceborne: Landsat-8, Private soil data	I4, I20	Model - BPNN; Results - R^2 0.947	1. Captured the nuanced relationship between plant health and SOM.	1. Exploration of limited spectral indices.
John et al. (2020) [326]	MS	Nigeria	SOC	Spaceborne: Landsat-8, Private soil data	I4, I10, I11, I12, I17, I18, I19, I20	Model - RF; Results - R^2 0.68	1. Explored SOC and spectral indices correlations.	1. Only one nutrient targeted.
Kumar et al. (2018) [327]	MS	India	SOC	Spaceborne: Aster, Private soil data	I4, I21, I22, I23, I24	Model - Regression Kriging; Results - RMSE 0.196	1. Explored SOC and spectral indices correlations.	1. Soil moisture and temperature are not considered. 2. Only one nutrient targeted.
Xu et al. (2018) [328]	MS	India	N	Spaceborne: Landsat-8, Private soil data	I4, I11, I13, I14, I20, I25, I26, I27, I28, I29	Model - Regression Kriging; Results - Adjusted R^2 0.56	1. Explored nitrogen and spectral indices correlations.	1. Only one nutrient targeted.
Kalambukattu et al. (2018) [63]	MS	India	SOC, N, P	Spaceborne: Landsat-8, Private soil data	I15, I16, I21, I30	Model - ANN; Results - R^2 0.83, 0.62, 0.511 for SOC, N, P	1. Explored soil nutrient and spectral indices correlations.	1. Soil moisture and temperature are not considered.
Jin et al. (2017) [148]	HS	China	SOM	Ground Based: Spectrometer, Private soil data	I4, I5, I15	Model - PSO-SVM; Results - R^2 0.43	1. Explored SOM and spectral indices correlations.	1. Only one nutrient is targeted.

Note: HS-Hyperspectral, MS-Multispectral, I1-Yellowness Index, I2-ferrous oxide index, I3-carbonate level, I4-NDVI, I5-EVI, I6-ARVI, I7-NPCRI, I8-RDVI, I9-SCI, I10-NDMI, I11-SAVI, I12-RVI, I13-NDWI, I14-NDSI, I15-RI, I16-CI, I17-Clay Index, I18-Iron Index, I19-NDBI, I20-LST, I21-VTCI, I22-BI, I23-WI, I24-GI, I25-NDGI, I26-GCI, I27-NDRE, I28-BSI, I29-MSI, I30-HI

a single nutrient or characteristic. Hence, this thesis aims to address these drawbacks by developing soil-specific spectral indices for accurate and comprehensive predictions of soil nutrient levels.

3.6 Limitations of Existing Machine Learning in Soil Nutrient Prediction

Recent advancements have been made in the study of soil nutrient prediction using spectral data. To increase the accuracy and robustness of the models, researchers must address several research gaps, which are as follows:

1. ***Limited availability of open ground truth datasets:*** The lack of widespread availability of high-quality, standardized datasets in different places poses difficulties in constructing dependable models. Utilizing comprehensive open datasets would improve the precision and practicality of soil nutrient prediction models [46, 55].

2. ***Limited coverage of hyperspectral sensors:*** The hyperspectral sensors acquire an extensive array of wavelengths to obtain precise spectral data, which is highly useful for predicting soil nutrients. Nevertheless, these sensors frequently possess restricted spatial coverage, which poses difficulties in expanding the application across vast regions without compromising resolution or data accessibility [287, 329].

3. ***Small sample size in ground truth dataset:*** The utilization of a limited sample size in a ground truth dataset can constrain the precision and dependability of soil nutrient prediction models. With a limited number of samples, the models may require assistance in accurately representing the diverse range of soil qualities found in different regions. For example, previous research, such as Zhang et al. [90], utilized a small sample size of 56 and achieved an R^2 value of just 0.58. Similarly, Pechanec et al. [206] employed a mere 22 soil samples and obtained an R^2 value of 0.44. As a result, the predictions made by the models are less reliable and less applicable to various scenarios.

4. ***Limited coverage of Deep Learning:*** . Despite the ability to capture complex, non-linear relationships in high-dimensional hyperspectral data, the application of DL models for soil nutrient prediction remains limited. Most existing studies mainly rely on ML methods, which may lack the advanced feature extraction capabilities of DL [86, 87]. Data availability and model complexity are challenges in using DL for soil nutrient prediction. While ML approaches perform reasonably well, DL models offer greater potential to exploit rich spectral information, yet remain underutilized in this domain.

5. ***Require the creation of hybrid models utilizing both ML and DL techniques:*** A hybrid technique combines the advantages of different ML and DL models, resulting in more reliable and accurate predictions of soil nutrient levels. Only a few studies have employed hybrid models for predicting soil nutrient levels [200, 302]. Hence, employing hybrid models to improve prediction accuracy and adaptability across various soil nutrients is necessary.

6. ***Need for the development of feature selection techniques to identify suitable wavelengths:*** Existing studies on feature selection for soil nutrient prediction have primarily relied on techniques such as PCA, discriminant analysis, recursive feature elimination, and correlation-based methods [90,91]. While these methods are commonly used, each approach has limitations — such as sensitivity to noise, difficulty in interpretation, and a bias toward capturing linear relationships. Therefore, there is a pressing need for a comprehensive comparative assessment of linear and nonlinear feature selection techniques. Such an evaluation would help identify the most effective approaches for selecting relevant spectral bands and wavelengths, enhancing the accuracy and interpretability of soil nutrient prediction models.

7. **Limited research on selecting the optimal spectral band subset for soil nutrient prediction using hyperspectral data:** Feature selection is a crucial step in hyperspectral data analysis for soil nutrient prediction, as it helps reduce data dimensionality and enhances model performance [85]. However, there remains a lack of comprehensive research focused on determining the optimal subset of spectral bands that best capture the most relevant information for

accurate nutrient prediction, which is crucial for enhancing model efficiency and scalability [90, 91].

8. **Underexplored development of novel spectral indices:** Although spectral indices have been widely used for vegetation monitoring, the development of tailored indices for predicting soil nutrients remains underexplored [30, 50]. A recent study by Gulhane et al. [30] utilized multispectral data and proposed novel indices for soil pH, iron, and phosphorus, highlighting the potential. However, similar advancements using hyperspectral data are still lacking, leaving a notable gap in current research.

3.7 Summary

This chapter first provides an overview of various publicly available multispectral and hyperspectral datasets commonly used in remote sensing and soil analysis. It outlines the key characteristics, including spectral resolution, spatial coverage, and sensor type, emphasizing the importance of supporting data-driven approaches for predicting soil nutrients. Building on this foundation, the chapter highlights the need to modernize soil nutrient analysis using spectral data and AI. It discusses how integrating hyperspectral/multispectral sensing with ML and DL enables faster and more accurate nutrient prediction. The chapter also mentions feature selection, model evaluation metrics, and spectral indices, which further enhance accuracy, robustness, and real-world applicability. It also identifies current limitations and open challenges in machine learning approaches. These advances support the achievement of the SDGs (SDG 2 – Zero Hunger, SDG 12 – Responsible Consumption and Production, SDG 13 – Climate Action, SDG 15 – Life on Land), enabling optimized fertilization, boosting yields, and promoting soil health, thereby opening new avenues for innovation in AI-driven soil analysis.

This chapter is based on the following publication:

Journal 1: S. Jain, D. Sethia and K.C. Tiwari." A Critical Systematic Review on Spectral-Based Soil Nutrient Prediction Using Machine Learning." **Environmental Monitoring and Assessment** 196, 699 (July 2024). **(SCIE, Impact Factor: 3.6,**

Publisher: Springer). Doi: https://doi.org/10.1007/s10661-024-12817-6

Chapter 4

HyPLS-GPR: A Hybrid Machine Learning Model for Soil Nutrient Prediction Using Multispectral Data

This chapter develops and evaluates machine learning models for large-scale soil nutrient prediction using Landsat-8 [159] multispectral data and LUCAS-2015 topsoil measurements [2]. To overcome the scalability limits of lab analysis and the low spectral depth of RGB imagery, this work integrates multispectral observations with ground-truth data for cost-effective nutrient prediction. This thesis presents the comparison of multiple ML models—PLSR, SVR, MLP, and GPR—and proposes a hybrid HyPLS-GPR model that combines linear and nonlinear learning. It evaluates the proposed HyPLS-GPR model on the LUCAS-2015 dataset [2] and the Rupnagar dataset [5]. This work aims to develop a scalable, accurate, and region-specific model for predicting key soil nutrients using multispectral data and ML models.

4.1 Motivation

Large-scale soil nutrient prediction is vital for crop productivity, but traditional lab methods are costly and spatially limited [55], and RGB imagery lacks sufficient spectral depth [54]. To overcome these limitations, remote sensing offers a scalable alternative through spectral observations. Multispectral satellites, such as Landsat-8 [159], capture

reflectance in the VIS, NIR, and SWIR bands, which is linked to soil nutrients. Prior studies using Landsat-8 [159] demonstrated its effectiveness in estimating soil nutrients across diverse agricultural landscapes, highlighting its potential for scalable soil monitoring [265, 311]. When combined with ground-truth data, the multispectral observations enable accurate, data-driven nutrient prediction using machine learning, providing a cost-effective approach to monitor soil health across large agricultural regions [72, 159].

The LUCAS topsoil survey [1] with Landsat-8 [159] multispectral images has become a key benchmark for soil nutrient mapping in Europe. Existing studies either target single nutrients or prioritize large-scale mapping over local accuracy, limiting the usefulness for region-specific agricultural planning [266, 271]. A detailed discussion of these studies is provided in section 4.2. This thesis addresses the gap by applying ML models to Landsat-8 multispectral images [159] and the LUCAS-2015 [1] topsoil survey for Belgium and Luxembourg, considering all soil nutrients, including pH in H_2O, pH in $CaCl_2$ OC, N, P, K, CEC, and $CaCO_3$. It predicts multiple properties to enable scalable, accurate, and region-specific assessments of soil nutrients for precision agriculture. The following are the main contributions of this work:

1. ***Proposal of hybrid machine learning model (HyPLS-GPR) for regional soil nutrient prediction:*** This thesis integrates multispectral Landsat-8 satellite imagery [159] with the LUCAS-2015 topsoil survey [2] to predict soil nutrients in Belgium and Luxembourg. It compares four machine learning models—PLSR, SVR, MLP, and GPR—and proposes a hybrid model (PLSR+GPR) called HyPLS-GPR, which combines linear feature extraction with nonlinear probabilistic learning to improve prediction accuracy for region-specific nutrient predictions.

2. ***Validation of proposed HyPLS-GPR model on Rupnagar dataset:*** This thesis initially trains HyPLS-GPR on the LUCAS-2015 dataset [2] and then validates it on the Rupanagar multispectral dataset [5] to assess its generalisation capability. Evaluating the model on independent data demonstrates its robustness and predictive accuracy for region-specific soil nutrient prediction.

Table 4.1: Comparison of previous studies based on Lucas topsoil database and multispectral data for soil nutrients prediction

Previous Work	Research Focusses	Lucas top soil Dataset Year	Algorithm	Results
Liu et al. (2023) [330]	Spectral clustering-based prediction of the SOC in the LUCAS soil database.	2015	PLSR	R^2 0.75
Zhong et al. (2021) [191]	Investigated the modeling capability of DCNN for soil attributes.	2012	DCNN	RMSE is 19.130, 0.971, 6.614, 0.326 and 24.526 for SOC, N , CEC, pH and CaCO3
Gouda et al. (2021) [331]	Soil texture prediction using Lucas topsoil data.	2015	MLP, CatBoost and LightGBM	CatBoost shows highest performance with R^2 0.79, 0.76 and 0.85 for sand, silt and clay predictions.
Pflugmacher et al. (2019) [332]	Mapping pan european land use land cover using Landsat-8 and Lucas top soil database.	2015	RF	Accuracy 75.1 %
Castaldi et al. (2019) [201]	SOC prediction using Lucas topsoil database and sentinel data.	2012	Exponential Regression model (ERM)	RPD 2.9
Ballabio et al. (2019) [266]	Soil pH in H2O and CaCl2, EC, CaCO3, N, K prediction using lucas topsoil and modis satellite data.	2012	GPR	RMSE for EC 11.02, N 1.97, pH in H2O 0.78, pH in CaCl2 0.68 K 121.89 and CaCO3 78.29
Ballabio et al. (2018) [51]	To look at the distribution of copper in the soils of 25 Member States of the European Union (EU), this study used the LUCAS topsoil survey.	2012	GPR	Explained deviance 66%
Castaldi et al. (2018) [271]	SOC prediction using lucas topsoil data and XDS Rapid content analyzer spectroradiometer for Belgium and Luxembourg Region.	2012	PLSR	RMSE 3.7 and RPD 2.54

4.2 Related Work

Numerous previous studies have made efforts to understand the complexity of spatial data interpretation and its applications in the context of soil analysis and remote sensing. The primary focus of this work is the analysis of Lucas's topsoil ground truth and multispectral data. Table 4.1 shows previous work using Lucas's topsoil database and multispectral data for estimating soil nutrients. It compares the target soil nutrients and modeling approaches, highlighting trends and gaps in earlier research. This summary helps understand the effectiveness of different methods and identifies areas where further improvement or exploration is needed.

Liu et al. [330] compared clustering strategies in soil modeling, finding that k-means clustering with PLSR had similar prediction accuracy as local PLSR but significantly faster runtime, supporting real-time SOC quantification. Zhong et al. [191] demonstrated that Deep Convolutional Neural Network (DCNN) predicts soil nutrients and texture based on a substantial spectral dataset from the LUCAS database, achieving accurate predictions for most properties and outperforming traditional methods. Pflugmacher et al. [332] combined Landsat-8 and LUCAS data, using spectral-temporal

metrics to map 12 land cover/use classes across Europe. According to Ballabio et al. [51], a study on copper distribution in EU soils revealed that significant factors influencing copper distribution include topsoil characteristics, land cover, climate, and agricultural techniques. Notably, vines exhibit the highest copper concentration due to specific soil characteristics and humid conditions.

Castaldi et al. [271] showed that accurate SOC prediction across various spectral libraries is made possible by standardized multivariate calibration utilizing soil samples from the LUCAS database without further field or lab effort. According to Castaldi et al. [201], accurate large-scale SOC mapping in North-Eastern Germany is possible without the need for fresh field missions as a result of Sentinel-2 images, optimized Normalized Burn Ratio thresholds, and SOC models utilizing the LUCAS spectral library. Ballabio et al. [266] presented a study on LUCAS-based topsoil property mapping for the European Union, emphasizing chemical characteristics by employing harmonized data and a GPR model.

While previous studies provide valuable insights, most work has focused on single nutrients or specific regions, highlighting the need for robust, scalable models. The proposed HyPLS-GPR model addresses this gap by enabling the accurate prediction of all region-specific soil nutrients.

4.3 Experimental Methodology

This work outlines the methodology for the HyPLS-GPR model in predicting soil nutrients. Figure 4.1 presents the research methodology, integrating Landsat-8 [159] multispectral imagery with LUCAS-2015 [2] topsoil data for Belgium and Luxembourg. Spectral reflectance from seven bands serves as independent features, while soil nutrients (N, P, K, OC, EC, $CaCO_3$, pH in H_2O, pH in $CaCl_2$) act as dependent variables. These inputs are combined to develop a HyPLS-GPR model for the accurate prediction of soil nutrients.

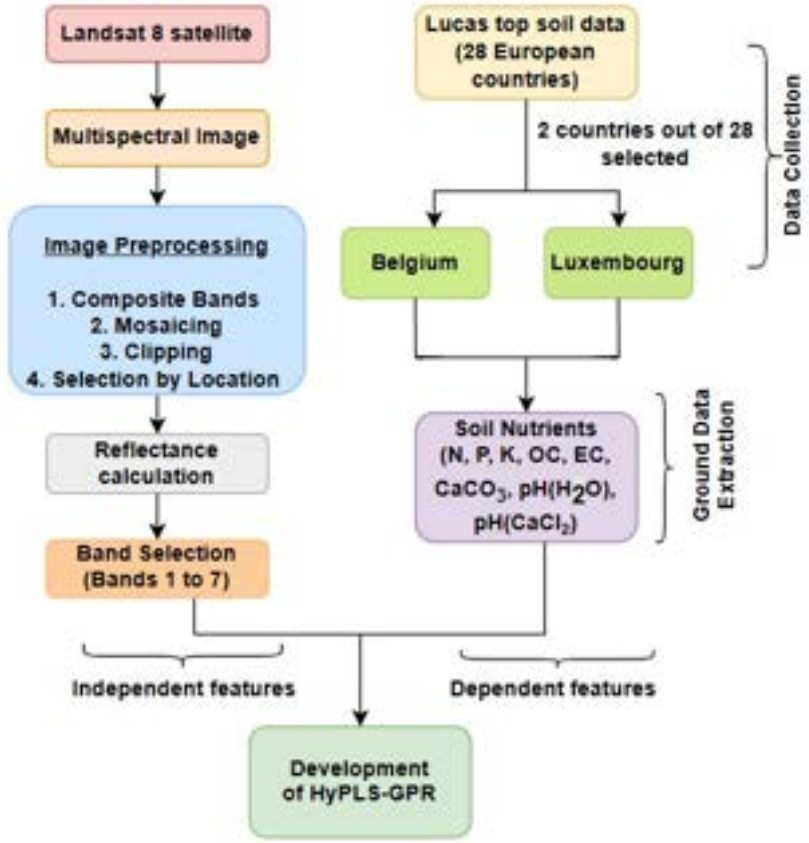

Figure 4.1: Research methodology flowchart for proposed HyPLS-GPR model

4.3.1 Dataset Description

This work utilizes an extensive database that includes the 2015 Lucas topsoil dataset [2] and the Rupnagar dataset [5], as described in chapter 3, along with Landsat-8,9 OLI/TIRS C2 L2 [159] satellite multispectral imagery. This thesis primarily focuses on LUCAS-2015 (Belgium and Luxembourg) to evaluate a wide range of soil nutrients, including N, P, K, OC, EC, $CaCO_3$, pH in H_2O, and pH in $CaCl_2$. At the same time, the Rupnagar region is used exclusively to validate the predictive performance of the proposed HyPLS-GPR for SOC and soil pH. The 2015 Lucas topsoil dataset covers 28 European nations and contains 92 ground-level soil measurements for Belgium and Luxembourg [333], while the Rupnagar dataset contains only 60 soil samples.

Figure 4.2 illustrates the key maps of Belgium, which display the spatial distribution of crucial soil characteristics, facilitating an understanding of the area's soil fertility, acidity, salinity, and nutrient content. The map depicts the spatial distribution of soil nutrient levels within the scope of interest. Each region's assigned color represents the

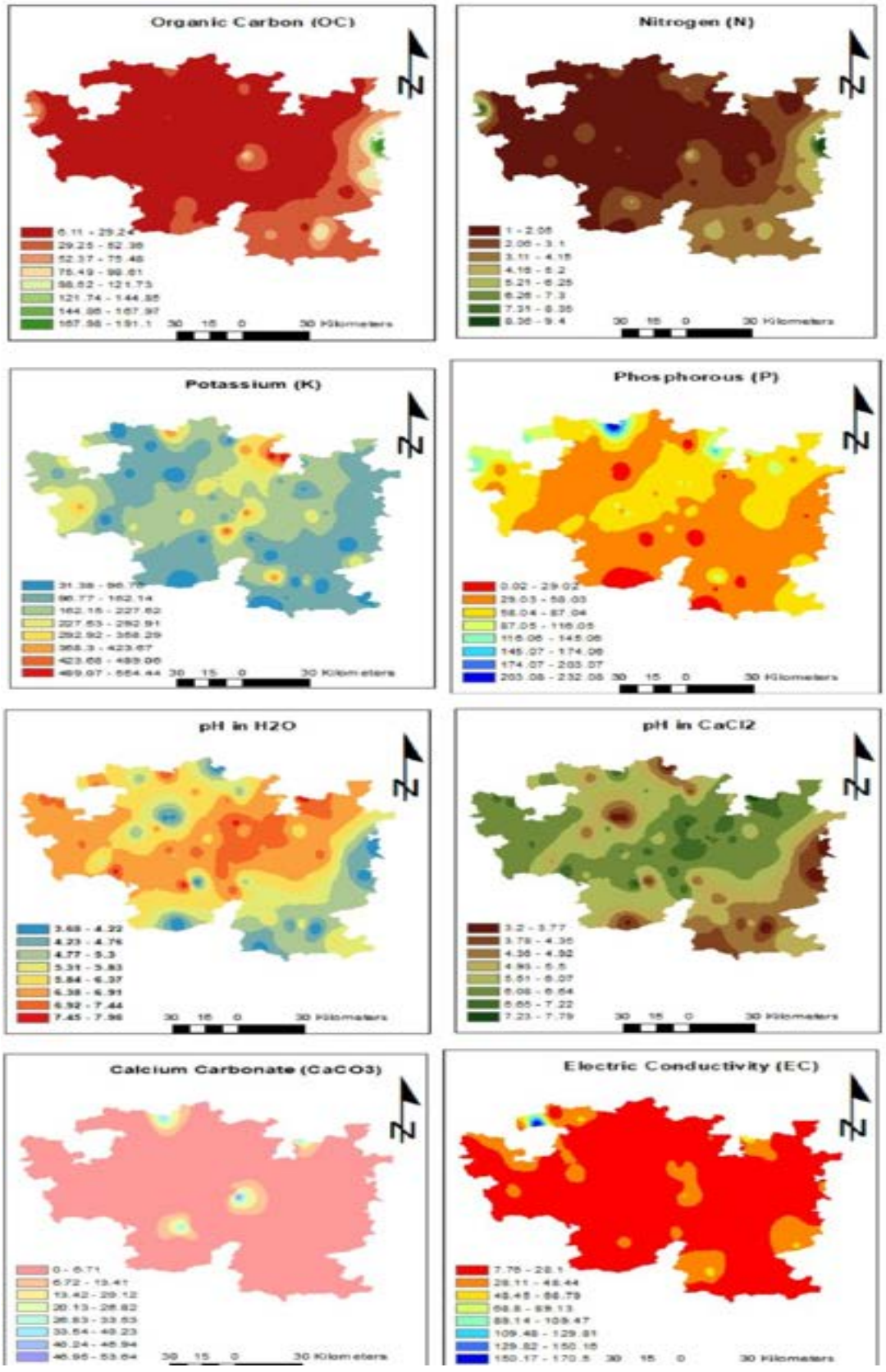

Figure 4.2: Top soil nutrients maps of Belgium using ArcMap/ArcGIS [19]

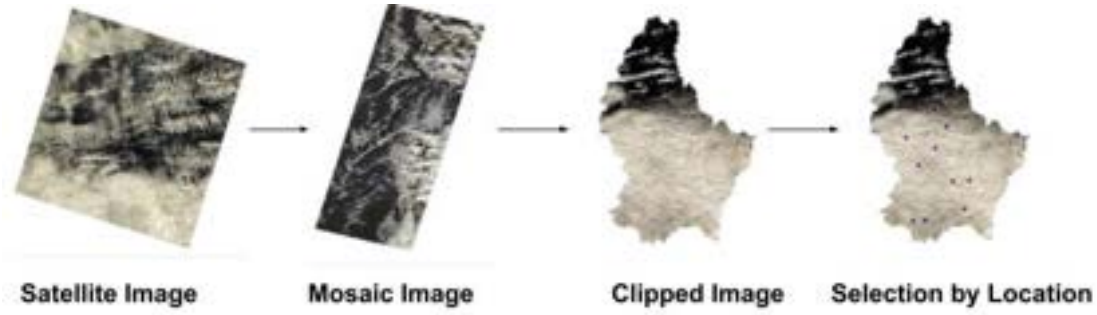

Figure 4.3: Image preprocessing steps of Landsat-8 multispectral images

range of soil nutrient concentrations, from minimum to maximum. Accordingly, lighter or darker tones indicate higher or lower concentrations of soil nutrients. A specific color denotes the presence of soil nutrients that fall within the stated range.

4.3.2 Data Preprocessing

Data Preprocessing is essential in realizing the full potential of Landsat-8,9 [159] OLI/TIRS C2 L2 satellite imagery for accurate soil nutrient prediction. This work describes the four significant preprocessing steps: band compositing, mosaicing, clipping, and location-based selection, as described in Chapter 2. Figure 4.3 illustrates the image preprocessing steps, showcasing the transformation of a multispectral image from a large region to a more confined target area for further investigation. These image preprocessing steps are performed using the ArcGIS tool [334], a geospatial software that enables viewing, editing, managing, and analyzing geographic data. Here are the simplified steps for performing these tasks:-

1. First, the ”Composite Bands” tool joins all seven bands.
2. Then, the ”Mosaic to New Raster” tool is used by setting the output location and the input rasters.
3. Then, the ”Clip” tool sets the clipping boundary for the Belgium and Luxembourg region.
4. At last, the ”Select by Location” tool is used by setting the target layer (LUCAS ground truth data) for selection and then picking the source layer (spectral data).

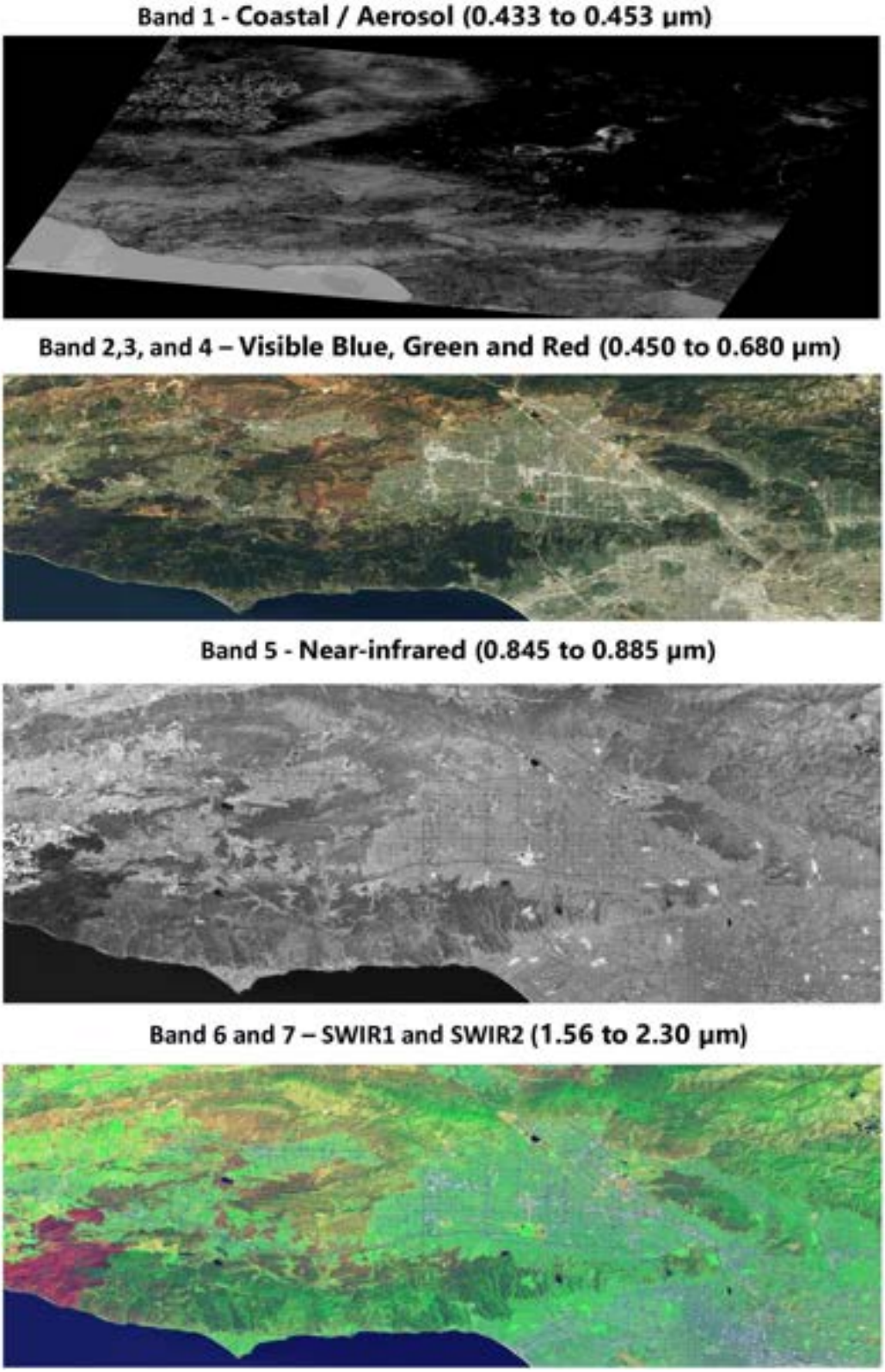

Figure 4.4: Landsat-8 bands [20]

4.3.2.1 Feature Engineering

This work reviews the process of generating various features and then normalizing them for consistency and simplicity. It ensures that all data inputs are on a comparable scale, thereby improving model performance and convergence.

1. ***Reflectance Calculation:*** An electromagnetic wave's reflection from a substance or surface having its reflectance spectrum is known as reflection [335]. This work calculates reflectance to connect the reflectance index with images of soil nutrients. An accurate measurement of reflected and transmitted light, including wavelengths invisible to the human eye, is crucial for predicting soil nutrient levels. The image acquisition involves utilizing the susceptible Landsat-8 OLI sensor, known for its exceptional sensitivity. Subsequently, the obtained data is converted into 16-bit DNs from 0 to 65536. The dataset is normalized using the *Standard Scaler*[1], which changes the features' means and standard deviations to 0 and 1, respectively [173]. This procedure ensures that scaling is uniform, improving algorithm performance by facilitating convergence, controlling outliers, and encouraging consistent feature behavior throughout the model.

2. ***Band Selection:*** Selecting influential and representative spectral bands is critical in eliminating redundant information and finding the highly correlated band with the targeted soil nutrient. Landsat-8,9 OLI/TIRS C2 L2 delivers seven imagery bands for each region [159]. Figure 4.4 presents Landsat-8 imagery across multiple spectral bands, illustrating how different wavelengths reveal distinct surface features. Coastal/aerosol Band 1 highlights water bodies and atmospheric aerosols, while Bands 2–4 (visible blue, green, and red) show natural-color landscapes, distinguishing vegetation, urban areas, and bare soil. Near-infrared Band 5 emphasizes vegetation health, and SWIR Bands 6–7, in a false-color composite, enhance differences in vegetation, soil moisture, and built-up areas [336]. The HyPLS-GPR model leverages all Landsat-8 [159] bands to harness multispectral data for land cover mapping, vegetation monitoring, and

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

environmental analysis.

3. ***Ground Data Extraction:*** The primary nutrients examined in this work are N, P, K, OC, EC, $CaCO_3$, and pH in H_2O, as well as pH in $CaCl_2$, which are crucial for enhancing soil fertility and crop productivity. The LUCAS topsoil dataset from 2015 served as a ground truth data source. The extraction of ground truth data from the LUCAS dataset encompasses two nations, Belgium and Luxembourg. The validation and enhancement of the soil nutrient prediction algorithm relied on this dataset as a reliable source of precise soil parameters within the specified regions.

4.3.3 Proposed HyPLS-GPR model

This work explains the design of the proposed HyPLS-GPR, a hybrid regression model. Initially, this thesis uses various ML models to predict soil nutrients. The HyPLS-GPR model outperforms standalone, independent, and traditional models.

4.3.3.1 Parameter Tuning and Regression Algorithms

For each algorithm, this thesis utilizes hyperparameter optimization to achieve optimal performance and efficiency in regression modeling. The Python library *sklearn.model_selection.GridSearchCV*[2] systematically explores parameter combinations to enhance forecast accuracy [337].

The dataset and model size are considered while selecting the hyperparameter grid. The *GridSearchCV* procedure keeps track of the hyperparameter combinations that performed the best on the validation data. The selected parameters are considered to be the best hyperparameters. *GridSearchCV* thoroughly analyzes multiple varieties of hyperparameters to find the ideal configuration for the model. Table 4.2 lists the hyperparameters utilized in this work for the various regression and suggested models. These parameters are carefully selected to optimize model performance and ensure robust predictions across different soil nutrient datasets. It also serves as a reference for reproducibility and comparison in future studies.

[2]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Table 4.2: Selected hyperparameters of regression models and HyPLS-GPR

Model	**Parameters**
PLSR	n_components=5, max_iter=50
SVR	kernel='rbf', C=0.5, gamma=0.15
MLP	activation_layer='relu', max_iter=15000, solver='sgd', learning_rate_init=0.05, hidden_layer_sizes=(3,2)
GPR	kernel='rbf', length_scale=9.0, normalize_y=False, random_state=1, aplha=18
HyPLS-GPR	n_components=5, kernel='rbf' n_restarts_optimizer=10

Table 4.3: Strengths and weaknesses of ML models and HyPLS-GPR

Model	**Strength**	**Weakness**
PLSR	1. Good for high-dimensional data with collinearity. 2. Handle both linear and nonlinear relationships.	1. Sensitive to the latent variables used.
SVR	1. Robust to outliers. 2. Implementation is easy.	1. It causes overfitting to small datasets. 2. It can be computationally expensive.
MLP	1. Flexible. 2. Handle complex relationships between variables.	1. It can be computationally expensive.
GPR	1. It can handle complex relationships between variables. 2. Good for interpolation.	1. It causes overfitting to small datasets. 2. It can be computationally expensive.
HyPLS-GPR	1. Improved accuracy. 2. Better interpretability. 3. Handles complex relationships.	1. It can be computationally expensive. 2. Complex implementation.

The examination encompasses several regression techniques employed to predict soil nutrient levels. These include PLSR, SVR, MLP, and GPR. Notably, compared to the other algorithms, PLSR and GPR demonstrated greater accuracy. Combining the advantages of PLSR and GPR, this thesis proposes a HyPLS-GPR model that leverages the complementary qualities to enhance soil nutrient prediction. Table 4.3 highlights the strengths and weaknesses of these ML models, and the proposed HyPLS-GPR model. It provides a comparative overview to show how each model performs in terms of accuracy, robustness, and computational efficiency. This table also helps to understand the advantages of the proposed approach over conventional models for predicting soil nutrients.

- *Partial Least Square Regression (PLSR):* It is a statistical method that builds latent components to detect linear correlations between input variables and a target variable [338].

- *Support Vector Regression (SVR):* It is an ML algorithm that finds a hyperplane

that optimizes the difference between projected values and actual data points to forecast a continuous target variable using support vector machines [339].

- *Multilayer Perceptron (MLP):* It is a type of artificial neural network that can recognize intricate patterns and connections in data due to its multiple layers of interconnected nodes [340].

- *Gaussian Process Regression (GPR):* It is a probabilistic ML technique that models data distributions and encapsulates prediction uncertainty [341].

- *HyPLS-GPR Model:* It utilizes both linear and probabilistic methodologies for improved predictive accuracy and uncertainty quantification. The HyPLS-GPR model combines the strengths of PLSR with GPR.

  Figure 4.5 shows the architecture of the HyPLS-GPR model for soil nutrient prediction. First, the model collects multispectral band reflectance values from the input spectral data and normalizes them using a standard scaler to ensure consistency. Then, it feeds the normalized features into different ML models (PLSR, SVR, MLP, and GPR), and then the top-performing ML models make a weighted averaging-based hybrid ensemble model with balanced weights, which combines predictions from PLSR and GPR. Finally, the ensemble generates predictions for multiple soil nutrients, including N, P, K, OC, EC, $CaCO_3$, pH in H_2O, and pH in $CaCl_2$.

4.3.4 Performace Metrics

This work utilizes the Root Mean Square Error (RMSE) and the Ratio of Performance to Deviation (RPD) to evaluate model performance, as described in chapter 3. Due to dataset-specific issues, R^2 is not considered a primary performance metric. In LUCAS-2015, despite moderate to high variance in several soil nutrients, the small (92) dataset can lead to unstable or hostile R^2 values. Even if the predictions are accurate, the R^2 values can become misleadingly low or unstable, giving a false impression of poor model performance [342]. Therefore, RMSE and RPD are the primary metrics, as these measures more reliably capture prediction accuracy and model performance for small-sized datasets [343].

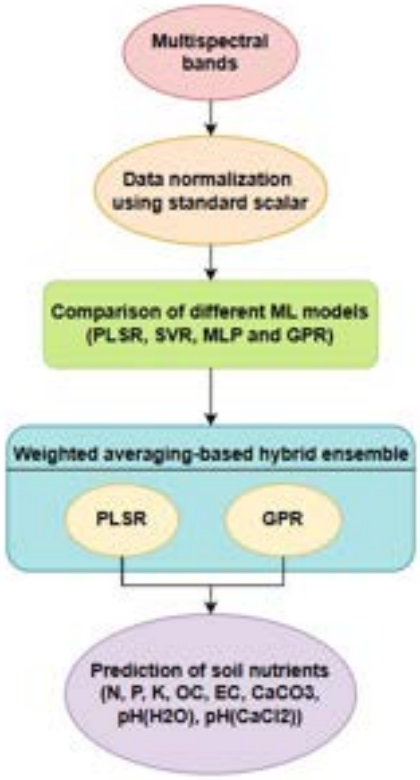

Figure 4.5: Architecture of proposed HyPLS-GPR model

4.4 Results and Discussion

In this chapter, four regression models, namely PLSR, SVR, GPR, and MLP, achieve the prediction of soil attributes, encompassing N, P, K, OC, EC, $CaCO_3$, pH in H_2O, pH in $CaCl_2$, and evaluating the accuracy and reliability of predictions involved using two crucial metrics: RMSE and RPD. Surprisingly, the results of this work reveal patterns in model performance across different soil conditions. Table 4.4 presents the results for various soil nutrients, including data from multiple regression models, such as the HyPLS-GPR model using LUCAS-2015 [2]. The table clearly compares model performance, highlighting that the proposed HyPLS-GPR model yields the most accurate predictions for each soil nutrient. For comparison, previously reported results from Castaldi et al. [271] are included, as the study also focuses on soil samples from Belgium and Luxembourg for SOC prediction, allowing a relevant benchmark to highlight the improvements achieved by the proposed hybrid approach.

Table 4.4: Comparison of ML models and HyPLS-GPR for soil nutrient prediction

SOC			**Nitrogen**		
Model	**RMSE**	**RPD**	**Model**	**RMSE**	**RPD**
PLSR	0.97	5.49	PLSR	1.11	3.92
SVR	1.04	5.19	SVR	1.1	4.06
MLP	0.95	4.57	MLP	1.07	3.98
GPR	0.99	5.36	GPR	1	4.34
HyPLS-GPR	**0.94**	**5.68**	**HyPLS-GPR**	**0.96**	**4.45**
Castaldi et al. (2018) (PLSR) [271]	3.7	2.54	-	-	-
Potassium			**Phosphorous**		
Model	**RMSE**	**RPD**	**Model**	**RMSE**	**RPD**
PLSR	1.11	3.91	PLSR	1.08	3.78
SVR	1.06	4.25	SVR	1.02	3.81
MLP	1.26	3.42	MLP	1.12	3.42
GPR	0.99	4.38	GPR	0.99	3.88
HyPLS-GPR	**0.98**	**4.5**	**HyPLS-GPR**	**0.98**	**3.95**
pH in H_2O			**pH in $CaCl_2$**		
Model	**RMSE**	**RPD**	**Model**	**RMSE**	**RPD**
PLSR	0.97	3.2	PLSR	0.98	3.37
SVR	1.05	2.92	SVR	1.07	3.06
MLP	0.97	3.06	MLP	0.97	3.25
GPR	0.98	3.17	GPR	0.99	3.33
HyPLS-GPR	**0.94**	**3.3**	**HyPLS-GPR**	**0.93**	**3.44**
$CaCO_3$			**EC**		
Model	**RMSE**	**RPD**	**Model**	**RMSE**	**RPD**
PLSR	1.09	3.67	PLSR	0.98	4.76
SVR	1.01	4.02	SVR	1.03	4.67
MLP	1.15	3.49	MLP	1.15	3.86
GPR	0.99	4.03	GPR	0.99	4.71
HyPLS-GPR	**0.92**	**4.34**	**HyPLS-GPR**	**0.97**	**4.81**

4.4.1 Individual Model Performance on LUCAS-2015 [2]

The predictive performance of the regression models is examined for each soil property in the LUCAS-2015 [2] to understand the strengths and limitations.

1. **SOC prediction:** PLSR and MLP produce the lowest RMSE (0.94–0.95) and highest RPD (4.57–5.68), demonstrating excellent predictive ability. GPR performs slightly comparable (RMSE 0.99, RPD 5.36), and SVR performs worst for SOC (RMSE 1.04, RPD 5.19).

2. **Nitrogen (N) prediction:** GPR achieves the lowest RMSE (1.00) and highest RPD (4.34), outperforming other models, while PLSR and SVR produce RMSE around 1.11–1.10 and RPD 3.92–4.06.

3. **Potassium (K) and Phosphorus (P) prediction:** GPR and SVR generally perform better for P and K (RMSE 0.99–1.06, RPD 3.81–4.38), while MLP showed moderate results.

4. **pH in H_2O and $CaCl_2$, Electrical Conductivity (EC):** PLSR consistently yields the lowest RMSE (0.94–0.97) and highest RPD (3.3–4.81), highlighting its strong linear modeling capability.

5. **Calcium Carbonate ($CaCO_3$):** GPR demonstrates superior performance (RMSE 0.99, RPD 4.03), indicating its effectiveness in capturing nonlinear patterns.

These results show that no single model dominates all soil attributes: PLSR performs best for SOC, pH, and EC; GPR excels in predicting N, P, K, and $CaCO_3$; MLP is competitive for SOC; SVR performs moderately across all parameters.

4.4.2 Rationale of the Proposed HyPLS-GPR Model

The HyPLS-GPR model is specifically designed using PLSR and GPR due to the complementary strengths and consistent performance across multiple soil nutrients. Analyzing individual models reveals that no individual regression model consistently outperforms others across all soil attributes. PLSR effectively captures linear relationships and performs best for variables such as SOC, pH, and EC. GPR effectively models nonlinear dependencies, showing superior results for N, P, K, and $CaCO_3$. Other models, such as SVR and MLP, demonstrate moderate performance but do not consistently achieve the lowest RMSE or highest RPD across all soil nutrients.

By combining PLSR and GPR using weighted averaging, the model leverages linear predictive power from PLSR and nonlinear flexibility from GPR, producing the lowest RMSE and highest RPD across all soil nutrients. Weighted averaging (PLSR-0.6 and GPR-0.4) allows the HyPLS-GPR model to adapt to the dominant pattern of each soil attribute, effectively functioning as a weighted averaging-based hybrid ensemble that

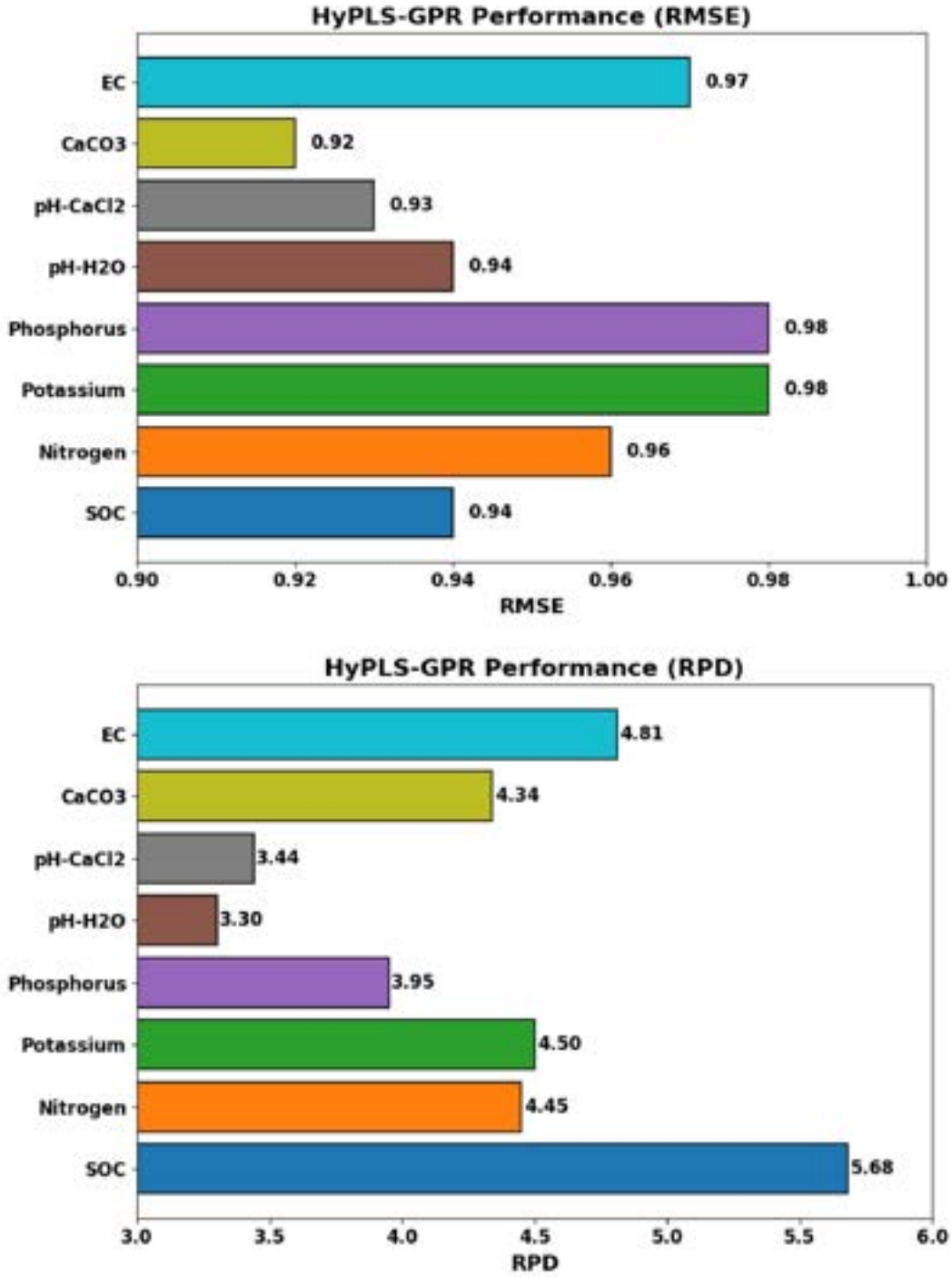

Figure 4.6: RMSE and RPD bar plots representing the performance of HyPLS-GPR

outperforms individual models across diverse conditions. The proposed HyPLS-GPR model achieves the lowest RMSE values, ranging from 0.92 to 0.98, and the highest RPD values, ranging from 3.3 to 5.68, outperforming both PLSR and GPR. These results highlight the robustness and general applicability of the hybrid approach. Figure 4.6 shows the performance of the HyPLS-GPR model across eight soil nutrients in terms

Table 4.5: Validation of HyPLS-GPR on Rupnagar dataset [5]

Soil Nutrient	Model	RMSE	RPD
SOC	PLSR	0.16	3.21
	SVR	0.16	3.36
	MLP	0.13	3.97
	GPR	0.14	3.78
	MLP (Tripathi et al. [5])	1.24	-
	HyPLS-GPR	**0.11**	**4.21**
Soil pH	PLSR	0.58	2.38
	SVR	0.53	2.74
	MLP	0.56	2.61
	GPR	0.75	1.94
	HyPLS-GPR	**0.46**	**2.97**

of RMSE and RPD. The results indicate consistently low RMSE values ($<$1.0) and high RPD values ($>$3.0), highlighting the model's strong predictive ability across all nutrients with RMSE values in the range of 0.92–0.98 and RPD values in the range of 3.30–5.68.

4.4.3 Validation of HyPLS-GPR on Rupnagar Dataset [5]

To test generalizability, the HyPLS-GPR model is validated on the Rupnagar multispectral dataset [5], which comprises 60 soil samples and 7 spectral bands, as described in chapter 3. In Rupnagar, the limited sample size (60 samples) combined with low variance in SOC: 0.02 and Soil pH: 0.27 makes R^2 highly sensitive and potentially misleading. Therefore, RMSE and RPD are used as primary evaluation metrics. Table 4.5 summarizes the predictive performance of different regression models for SOC and soil pH on the Rupnagar dataset [5], highlighting the superior performance of the HyPLS-GPR model. The key observations for SOC predictions are that the HyPLS-GPR model achieves the lowest RMSE (0.11) and highest RPD (4.21), outperforming all individual models. For soil pH predictions, the HyPLS-GPR model again reduces RMSE to 0.46 and increases RPD to 2.97. The previous study by Tripathi et al. [5] utilized the same Rupnagar dataset, but with different multispectral imagery (Sentinel-2 [238]), and reported a SOC RMSE of 1.24. HyPLS-GPR substantially improves accuracy, even with different multispectral satellite inputs, small sample sizes, and low-variance data.

4.4.4 Discussion

1. ***HyPLS-GPR Model Advantage:*** The proposed HyPLS-GPR model consistently outperforms individual models by capturing both linear (PLSR) and nonlinear (GPR) relationships, resulting in lower RMSE and higher RPD across all soil nutrients. This combination ensures accurate prediction for soil nutrient attributes even under varying spectral conditions.

2. ***Generalizability:*** Validation on the Rupnagar dataset, with only 60 soil samples, confirms that the HyPLS-GPR model performs robustly even with small datasets and low variance. The approach can be applied to other multispectral soil datasets with limited samples.

3. ***Regional Soil Prediction:*** The HyPLS-GPR provides a reliable regional soil nutrient prediction framework, supporting precision agriculture and environmental monitoring. It enables site-specific soil management decisions by accurately estimating multiple soil nutrients simultaneously.

4. ***Weighted Averaging-Based Hybrid Ensemble Justification:*** The weighted combination of PLSR (0.6) and GPR (0.4) reduces individual model weaknesses, effectively balancing linear and nonlinear predictive strengths, and enhancing overall accuracy. This ensemble approach ensures the HyPLS-GPR model consistently outperforms single models across all soil attributes.

4.5 Summary

This chapter utilizes multispectral satellite data to evaluate machine learning approaches for predicting soil nutrients. It compares four ML models—PLSR, SVR, MLP, and GPR—highlighting the strengths in handling spectral collinearity, nonlinearity, complex interactions, and uncertainty. Building on this comparison, it proposes a HyPLS-GPR model that integrates linear latent variable extraction with probabilistic nonlinear learning, improving prediction accuracy and robustness. This thesis demonstrates the potential of multispectral data and the HyPLS-GPR model for scalable, region-specific assessments of soil nutrients in Belgium and Luxembourg.

This chapter is based on the following publication:

Conference 2: S. Jain, D. Sethia and K.C. Tiwari. "A Hybrid Approach for Soil Nutrient Estimation using Multispectral Data in Belgium and Luxembourg." **in IEEE *Proc*. First International Conference on Advances in Electrical, Electronics, and Computational Intelligence (ICAEECI), October 2023.**

Chapter 5

CNViS-Net: A Self-Attention Framework for Soil Nutrient Prediction Using Hyperspectral Data

The previous chapter 4 explores machine learning models with multispectral data, which are limited by a few spectral bands, restricting the ability to capture complex, nonlinear soil spectral patterns. In contrast, deep learning on hyperspectral data leverages its rich, continuous spectra to detect subtle variations in soil. To harness these advantages, this thesis introduces CNViS-Net, a novel self-attention-based framework that integrates Convolutional Neural Networks (CNN) for compact spectral feature extraction and Vision Transformers (ViT) for capturing global dependencies. It is trained on the HYPERVIEW dataset [4] and validated on the Univalle dataset [6, 18], thereby enhancing generalization capability and computational efficiency.

5.1 Motivation

Multispectral imaging has been commonly used in soil studies due to its simplicity and lower data volume, but it provides only limited spectral information across a few discrete bands [15, 81]. Hyperspectral imaging, on the other hand, captures continuous spectral data across hundreds of narrow bands, offering much finer spectral resolution for soil nutrient prediction [147]. It enables the detection of subtle differences in

soil composition that multispectral data often misses [36]. However, hyperspectral data's high dimensionality and complex data structure present challenges for traditional ML models [344, 345]. Manual feature extraction can overlook essential patterns, and standard models, such as RF, SVM, and CNN, may struggle with spectral-spatial dependencies and computational efficiency.

Recent studies highlight the potential of Vision Transformers (ViTs) in hyperspectral analysis [346]. ViTs effectively capture local and global spectral features, achieving strong results in urban ground object classification [293] and soil texture classification [347]. Yet, the application to soil nutrient prediction remains largely unexplored. This thesis addresses this gap using the HYPERVIEW dataset [4], which contains high-dimensional hyperspectral soil data. Prior work on this dataset relied mainly on CNNs and traditional models [17,192,195], often underutilizing complex spectral interactions (as described in section 5.2).

To advance the state of the art, this thesis proposes a ViT-based framework for predicting key soil nutrients—Phosphorus (P), Potassium (K), Magnesium (Mg), and pH—aiming to improve both accuracy and computational efficiency in hyperspectral modeling for precision agriculture.

The main contributions of this work are as follows:

1. ***Proposal of a novel self-attention framework - CNViS-Net:*** The framework predicts soil nutrients by integrating advanced feature extraction techniques like CNN and ViT with an SVR model using HYPERVIEW hyperspectral dataset [4]. The method increases the prediction accuracy and computational efficiency of soil nutrient prediction.

2. ***Validation of proposed CNViS-Net framework on Univalle dataset:*** This work tests the proposed CNViS-Net framework on the HYPERVIEW dataset [4] and subsequently validates it on the Univalle dataset [6] to assess its generalisation performance. This validation on an independent dataset highlights the model's robustness and adaptability in predicting soil nutrients across different soil types and conditions, demonstrating its effectiveness for soil analysis.

5.2 Related Work

The prediction of soil nutrients using hyperspectral data has garnered significant attention in recent years, leading to the development of various innovative methodologies. Table 5.1 summarizes the previous work on the HYPERVIEW dataset [4]. It details the target soil nutrients, modeling techniques, and performance metrics reported in earlier studies. This overview helps identify research trends and gaps, guiding the development of improved models for predicting soil nutrients.

Kuzu et al. [195] combined feature extraction techniques like Wavelet and Fast Fourier Transform with ML models, achieving good results for P, K, Mg, and pH. Still, the reliance on manual feature engineering can be more efficient, but it may also miss important spectral information. Zelenka et al. [17] utilized SpectralNet3D with input standardization, achieving good results for all nutrients but exhibiting inconsistent performance for phosphorus (0.976 Score), suggesting a lack of generalization across different soil conditions. Gacek et al. [192] applied the Savitzky-Golay filter, obtaining strong results, but using complex DL architectures raises concerns about overfitting and interpretability.

The use of ViTs in hyperspectral image analysis has shown promising results across various applications. For instance, Wang et al. [293] achieved high classification accuracy with ViT across multiple datasets, including WHU-Hi-LongKou, Pavia University, Xuzhou, and Indian Pines, with accuracies of 99.89%, 99.76%, 99.80%, and 98.81%, respectively. It highlights the potential of ViT in advancing hyperspectral image classification. Similarly, Kuhnlein et al. [347] utilized CNN and ViT models to classify soil textures using the LUCAS 2009 and 2015 datasets, achieving an accuracy of 67.62%, which further showcases ViT's capability in soil-related hyperspectral data. Despite these successes, there remains considerable scope for applying ViT in soil property analysis, especially for predicting nutrient levels and other soil characteristics. While current research largely focuses on classification, extending ViT applications to regression tasks in soil analysis can unlock valuable insights.

In summary, the previous studies on HYPERVIEW dataset [4] highlight the challenges associated with manual feature extraction, generalization across nutrients, and model complexity. While previous studies on ViT have shown significant potential

Table 5.1: Previous work on HYPERVIEW dataset [4]

Reference	Feature extraction technique	Model	Score
Kuzu et al. [195]	All techniques combined - wavelet transform, fast fourier transform, singular value decomposition, first, second and third derivative of the original data	Hybrid regressor (RF+KNN)	P, K, Mg, pH 0.855, 0.807, 0.725, 0.749
Zelenka et al. [17]	Input-standardization	SpectralNet3D	P, K, Mg, pH 0.976, 0.439, 0.814, 0.889
Gacek et al. [192]	Savitzky Golay	SVR, RF, CNN2D	CNN2D for P, K, Mg 0.92 0.98 0.88 and SVR, RF for pH 0.93

Score metric used by the Hyperview challenge [195] is defined in Chapter 3.

in hyperspectral data analysis, the application in soil nutrient prediction is limited, with most studies primarily focusing on classification tasks rather than quantitative predictions. Hence, this thesis introduces a novel framework that overcomes these limitations by integrating advanced feature extraction methods, such as CNNs and ViTs, which enable the automatic learning of spectral patterns and improve prediction accuracy and computational efficiency for soil nutrient prediction.

5.3 Experimental Methodology

5.3.1 Dataset Description

The HYPERVIEW dataset [4], provided by the HYPERVIEW Challenge organizers, consists of 2,886 hyperspectral images of soil patches collected from agricultural areas in Poland. Among these images, a subset of 1,732 patches is paired with corresponding soil parameters, including P, K, Mg, and pH. The organizers reserved the ground truth for the remaining 1,154 patches for the challenge's final ranking. In addition, the Univalle dataset [6], comprising 152 soil samples with a wide range of phosphorus concentrations, has been used to validate the generalization performance of the proposed model on an independent dataset. A detailed description of these datasets is provided in Chapter 3.

5.3.2 Feature Extraction

In this work, feature extraction focuses solely on spectral features from hyperspectral data. The process begins with applying a CNN as a compression tool to reduce the spectral dimensions of the input data. The CNN block consists of convolutional layers that progressively compress the number of channels while retaining essential spectral information. It achieves this by applying convolutional layers, where the input hyperspectral data $\mathbf{X} \in \mathbb{R}^{B \times H \times W}$ is convolved with filters W_i to extract feature maps F_i as computed in Equation 5.1.

$$F_i = \sigma(X * W_i + b_i) \tag{5.1}$$

where σ is a non-linear activation function, and b_i is the bias term. These compressed feature maps pass to the ViT, which treats the compressed outputs as tokens and applies self-attention to capture long-range dependencies between spectral features. This work utilizes a pre-trained ViT model (ViT-Base-Patch16-224) from the Hugging Face library for feature extraction, leveraging its robust representation learning capabilities. The ViT operates by splitting the input feature maps into patches, embedding them, and processing them through multiple transformer layers. The combination of the patch embeddings and the positional embeddings, forming the input sequence for the transformer layers in the ViT model (Z_0), is computed in Equation 5.2.

$$Z_0 = [X_{\text{patch1}}E;\ X_{\text{patch2}}E;\ \ldots;\ X_{\text{patch}N}E] + E_{\text{pos}} \tag{5.2}$$

where E is the embedding matrix, and E_{pos} adds positional information. The transformer layers capture spectral correlations, and regression models use the output for further predictions.

5.3.3 Architecture of Proposed CNViS-Net Framework

The proposed self-attention framework, CNViS-Net, integrates CNN, ViT, and SVR to predict soil nutrients from hyperspectral data. CNNs extract localized features, while ViTs capture long-range spectral relationships. The architecture processes compressed hyperspectral images (11, 11, 3) through four convolutional layers (kernel: 3×3,

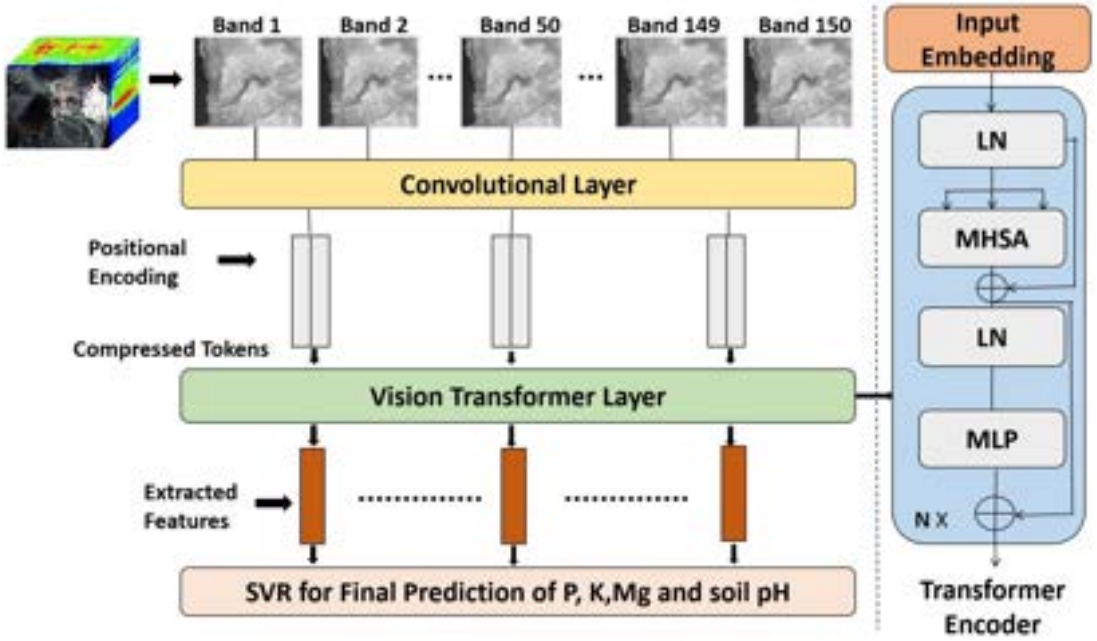

Figure 5.1: Architecture of the proposed CNViS-Net framework.

filters: 32, 32, 16, 3) with ReLU activation, followed by two max-pooling layers (2×2) for feature reduction. The ViT component enhances contextual learning with 12 transformer layers and 12 attention heads (hidden size: 768). Optimized using the Adam algorithm (learning rate: 0.0001), the model is trained with a batch size of 32 for 20 epochs. SVR (C=1.0) serves as the final regression step. This architecture combines the local feature extraction capabilities of CNNs and the global attention of ViTs, thereby improving the accuracy of soil nutrient prediction.

Figure 5.1 shows the detailed structure of the proposed CNViS-Net framework. The transformer encoder uses Layer Normalization (LN) before the MLP and Multi-Head Self-Attention (MHSA) functions. Residual connections are added to each self-attention and MLP layer to enhance trainability and fully utilize the information extracted at different layers. The input hyperspectral patches extend along the spectral dimension, and a convolutional layer embeds the features into three channels. Position encoding maintains the positional relationships between tokens after appending the output token to the input sequence. Finally, the extracted features pass into the SVR model to predict P, K, Mg, and soil pH. Table 5.2 shows the hyperparameters of the proposed CNViS-Net framework. The selection of these hyperparameters involved careful empirical testing. The process identified hyperparameters that consistently

Table 5.2: Hyperparameters outline the design of the CNViS-Net

Hyperparameters	**Values**
Input Shape (CNN)	(11, 11, 3)
Number of Classes	4
Convolutional Layers	4
CNN Kernel Size	(3, 3)
CNN Filters	32, 32, 16, 3
CNN Activation	ReLU
Max Pooling Layers	2
Max Pooling Size	(2, 2)
Vision Transformer Layers	12
ViT Patch Size	(16, 16)
ViT Hidden Size	768
ViT Attention Heads	12
ViT Learning Rate	0.0001
SVR Regularization Parameter	1.0
Batch Size	32
Epochs	20
Loss Function	Mean Squared Error
Optimizer	Adam
Metrics	RMSE, MAE, MSE, and Score

Table 5.3: Selected parameters of ML models combined with CNN and Transformer

Model	**Selected Parameters**
KNN	n_neighbors=5, weights='distance', algorithm='auto', leaf_size=50, p=2, metric='minkowski', metric_params=None, n_jobs=None
RF	n_estimators=200
PLSR	n_components=100,scale=False,max_iter=50
SVR	kernel='rbf', C=100, gamma=0.1, epsilon=.1

deliver the best results across key performance metrics, including RMSE, MAE, and MSE, for soil nutrient prediction using hyperspectral data.

5.3.4 Parameter Tuning and Regression Algorithms

This work utilizes the *sklearn.model_selection.GridSearchCV* [1] utility from the scikit-learn Python package to perform systematic hyperparameter tuning [337]. *GridSearchCV* evaluates combinations of hyperparameters by exhaustively searching

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

through a predefined parameter grid for each ML model. This approach ensures the identification of the optimal parameters that maximize the model's performance across various metrics. Table 5.3 presents the final hyperparameters selected for various ML models, showcasing the configurations that achieved the best predictive accuracy and generalizability for soil nutrient prediction using hyperspectral data. These selected parameters are a reference for replicating the experiments and guiding future studies in similar spectral-based soil prediction tasks. The regression models utilized in this work include K-Nearest Neighbors (KNN), which predicts the target variable by averaging the values of the K closest data points. RF is an ensemble method that builds multiple decision trees and averages the predictions. PLSR projects predictors and the response variable into a new space for linear regression. SVR minimizes model complexity while maintaining a tolerance margin from actual target values. A comparative analysis with baseline models further highlights the improvements achieved by CNViS-Net.

5.3.5 Performance Metrics

This work evaluates the framework's performance across different soil nutrient predictions by analyzing key indicators like RMSE, MAE, MSE (as described in chapter 3), and score metric. The score metric, as discussed in chapter 3, accounts for the improvement over the baseline model provided by the HYPERVIEW challenge, which predicts the average value of each soil nutrient (MSEbl). A lower score is better.

$$\text{Score} = \frac{\text{MSE}_{\text{algo}}}{\text{MSE}_{\text{bl}}} \tag{5.3}$$

5.4 Results and Discussion

This work evaluates the effectiveness of the proposed CNViS-Net framework using various performance metrics like RMSE, MAE, MSE, and Score. Tables 5.4, 5.5, 5.6, and 5.7 show the proposed CNViS-Net framework's superior performance across all four soil nutrient predictions— P, K, Mg, and soil pH, respectively—when compared to both base models and state-of-the-art approaches. The results demonstrate the effectiveness of the CNViS-Net framework in capturing complex spectral patterns and

enhancing prediction accuracy compared to existing methods.

5.4.1 Individual Model Performance on HYPERVIEW [4]

The predictive performance of the regression models is examined for each soil nutrient for the HYPERVIEW dataset [4] to understand the strengths and limitations.

1. ***Phosphorus prediction:*** Compared to traditional hybrid approaches such as Kuzu et al. [195] (Score = 0.855), Zelenka et al. [17] (0.976), and Gacek et al. [192] (0.92), CNViS-Net achieves a substantially lower score, demonstrating improvement in the accuracy and robustness. Within the baseline ML comparisons, CNN+ViT with KNN (RMSE = 12.55, Score = 0.181) and RF (RMSE = 12.46, Score = 0.177) are competitive but still inferior to CNViS-Net. CNN+ViT with PLSR performs poorly (RMSE = 37.68, Score = 1.63), highlighting the limitations of linear regression in capturing the nonlinear relationships between soil and spectra.

2. ***Potassium prediction:*** Compared to prior state-of-the-art methods (Kuzu et al. [195], Zelenka et al. [17], and Gacek et al. [192]), which reported scores ranging between 0.439 and 0.98, CNViS-Net produces a substantially lower score, emphasizing its improved predictive capability. CNN+ViT with RF (RMSE = 17.55, Score = 0.08) and KNN (RMSE = 19.95, Score = 0.103) are competitive but marginally less accurate. PLSR underperforms, with an RMSE of 48.6 and a score of 0.3262, reaffirming its inadequacy for nonlinear spectral–nutrient mapping. Thus, CNViS-Net demonstrates a clear superiority for potassium prediction, outperforming prior studies and baseline ML models.

3. ***Magnesium prediction:*** Previous studies such as Kuzu et al. [195] (score = 0.725), Zelenka et al. [17] (score = 0.814), and Gacek et al. [192] (score = 0.88) reported notably higher score values, indicating inferior performance relative to CNViS-Net. Among the ML baselines, CNN+ViT with KNN (RMSE = 27.25, score = 0.47) and RF (RMSE = 28.02, score = 0.502) are less effective, while PLSR (RMSE = 62.8, score = 2.56) again performs poorly. CNViS-Net achieves the lowest RMSE and error scores, confirming its robustness for magnesium

Table 5.4: Comparison of CNViS-NET for phosphorus (P) prediction

Reference	**Feature Extraction Technique and Model**	**RMSE**	**MAE**	**MSE**	**Score**
Kuzu et al. [195]	Combined Wavelet Transform, Fast Fourier Transform, Singular Value Decomposition, First, Second and third derivative of the original data with Hybrid model (RF+KNN)	-	-	-	0.855
Zelenka et al. [17]	Input standardization SpectralNet3D	-	-	-	0.976
Gacek et al. [192]	Savitzky Golay CNN2D	-	-	-	0.92
Proposed framework	CNN and ViT with KNN	12.55	9.87	157.5	0.181
	CNN and ViT with RF	12.46	10.47	155.31	0.17
	CNN and ViT with PLSR	37.68	33.62	1420.35	1.63
	CNN and ViT with SVR (Proposed CNViS-Net)	**11.94**	**9.56**	**142.59**	**0.164**

Table 5.5: Comparison of CNViS-NET for potassium (K) prediction

Reference	**Model**	**RMSE**	**MAE**	**MSE**	**Score**
Kuzu et al. [195]	Combined Wavelet Transform, Fast Fourier Transform, Singular Value Decomposition, First, Second and third derivative of the original data with Hybrid model (RF+KNN)	-	-	-	0.807
Zelenka et al. [17]	Input-standardization withSpectralNet3D	-	-	-	0.439
Gacek et al. [192]	Savitzky Golay CNN2D	-	-	-	0.98
Proposed framework	CNN and ViT with KNN	19.95	16.55	398.34	0.103
	CNN and ViT with RF	17.55	14.93	308	0.08
	CNN and ViT with PLSR	48.6	40.91	2362.33	0.61
	CNN and ViT with SVR (Proposed CNViS-Net)	**16.88**	**15.4**	**285**	**0.074**

prediction.

4. ***Soil pH Prediction:*** Prior studies reported higher scores: Kuzu et al. [195] (0.749), Zelenka et al. [17] (0.889), and Gacek et al. [192] (0.93), reflecting reduced performance compared to CNViS-Net. Baseline CNN+ViT hybrids with RF (RMSE = 0.24, Score = 0.62) and KNN (RMSE = 0.26, Score = 1.16) are close competitors but slightly inferior to CNViS-Net. The PLSR variant (RMSE

Table 5.6: Comparison of CNViS-NET for magnesium (Mg) prediction

Reference	Model	RMSE	MAE	MSE	Score
Kuzu et al. [195]	Combined Wavelet Transform, Fast Fourier Transform, Singular Value Decomposition, First, Second and third derivative of the original data with Hybrid model (RF+KNN)	-	-	-	0.725
Zelenka et al. [17]	Input-standardization withSpectralNet3D	-	-	-	0.814
Gacek et al. [192]	Savitzky Golay CNN2D	-	-	-	0.88
Proposed framework	CNN and ViT with KNN	27.25	23.23	742.83	0.47
	CNN and ViT with RF	28.02	25.71	785.35	0.502
	CNN and ViT with PLSR	62.8	57.39	3944.54	2.56
	CNN and ViT with SVR (Proposed CNViS-Net)	**21.91**	**0.76**	**480.17**	**0.307**

Table 5.7: Comparison of CNViS-NET for soil pH prediction

Reference	Model	RMSE	MAE	MSE	Score
Kuzu et al. [195]	Combined Wavelet Transform, Fast Fourier Transform, Singular Value Decomposition, First, Second and third derivative of the original data with Hybrid model (RF+KNN)	-	-	-	0.749
Zelenka et al. [17]	Input-standardization withSpectralNet3D	-	-	-	0.889
Gacek et al. [192]	Savitzky Golay SVR and RF	-	-	-	0.93
Proposed framework	CNN and ViT with KNN	0.26	0.25	0.07	1.16
	CNN and ViT with RF	0.24	0.22	0.05	0.83
	CNN and ViT with PLSR	0.36	0.32	0.13	2.16
	CNN and ViT with SVR (Proposed CNViS-Net)	**0.23**	**0.2**	**0.04**	**0.66**

= 0.36, Score = 1.02) is again suboptimal. Overall, CNViS-Net outperforms all existing methods in terms of accuracy and consistency for soil pH prediction.

Across all four nutrients (P, K, Mg, and pH), the proposed CNViS-Net consistently outperforms traditional preprocessing-based models (Wavelet + FFT + SVD + RF/KNN, SpectralNet3D, Savitzky–Golay). Also, the combination of CNN+ViT with SVR outperforms the other classical ML regressors. PLSR consistently fails across all nutrients, confirming its limitation in handling nonlinear soil spectral

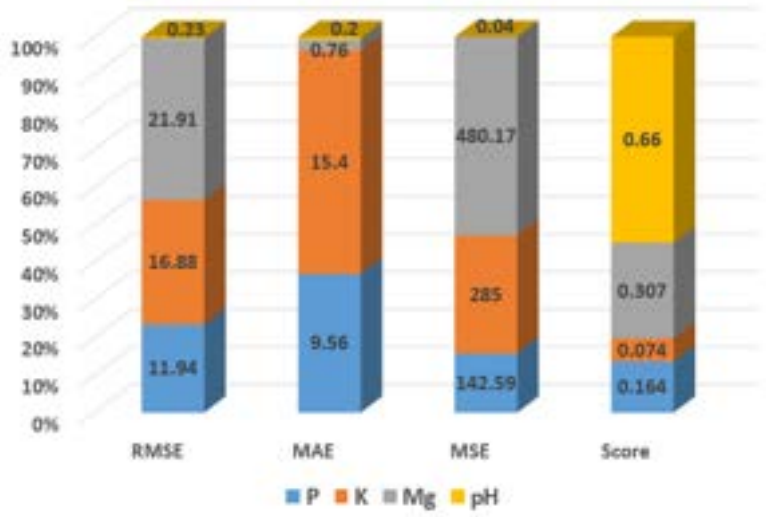

Figure 5.2: Performance metrics bar plots of the proposed CNViS-Net framework

relationships. RF and KNN perform reasonably well but remain slightly behind CNViS-Net, demonstrating the benefit of incorporating SVR within the CNN+ViT hybrid framework. The improvement margin is most evident for phosphorus and potassium, where CNViS-Net reduces RMSE and scores significantly relative to state-of-the-art and baseline models. Although other models perform competitively for pH prediction, CNViS-Net achieves the lowest error and Score, highlighting its generalizability across nutrient types.

Figure 5.2 presents a bar plot for performance analysis of the CNViS-Net framework in predicting soil nutrients. The results highlight that soil pH prediction achieves the lowest error, whereas Mg exhibits the highest RMSE and MSE, indicating variation in predictive accuracy across different nutrients. This visualization enables quick comparison of model performance across nutrients, highlighting areas where further model refinement may be necessary. Figure 5.3 displays the four scatter plots, which illustrate the measured versus predicted values for P, K, Mg, and soil pH, including both training and testing data points. Both the training (blue) and testing (yellow) values align closely with the perfect prediction line (green dashed), indicating highly accurate predictions across all nutrients. There is minimal deviation, showing that the proposed framework generalizes well and performs consistently on both datasets. It suggests that the proposed framework effectively captures the relationships between measured and predicted values for these soil nutrients.

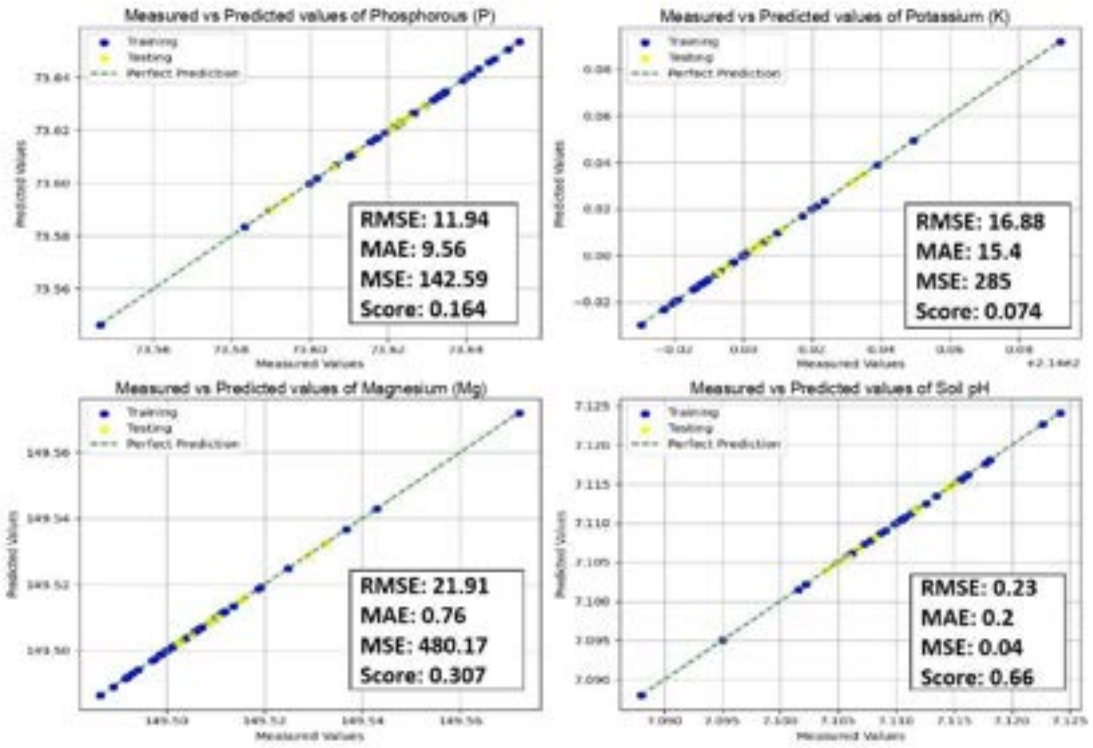

Figure 5.3: Scatter plots of the predicted values Vs. measured values of the proposed CNViS-Net framework

5.4.2 Rationale of the Proposed CNViS-Net Framework

- ***Effective feature extraction:*** The proposed CNViS-Net framework integrates CNN and ViTs for efficient feature extraction. This combination leverages the strengths of CNNs in capturing local spatial patterns and the global context capabilities of ViTs, leading to a more comprehensive understanding of soil spectral data. The framework enhances prediction accuracy by minimizing redundancy while preserving key spectral properties. By ensuring a more thorough examination of hyperspectral data, this hybrid approach enhances the calculation of soil nutrients.

- ***Superior regression performance:*** The use of SVR in combination with CNN and ViT ensures high prediction accuracy for multiple soil nutrients. It is especially well-suited for hyperspectral analysis because of its proficiency with high-dimensional data. Additionally, by reducing noise and improving consistency across multiple soil samples, the SVR component facilitates the refinement of the final forecasts.

Table 5.8: Validation of CNViS-Net on Univalle dataset [6] for soil phosphorus prediction

Approach	RMSE	MAE
ANN (Bolanos et al. [18])	22,991.80	-
CNN and ViT with KNN	22,020.45	17,842.30
CNN and ViT with RF	21,690.13	17,203.75
CNN and ViT with PLSR	21,740.67	17,510.92
CNN and ViT with SVR (Proposed CNViS-Net)	**21,507.30**	**16,984.60**

- ***Performance comparison of proposed framework:*** The comparison of the proposed CNViS-Net framework with the best state-of-the-art techniques shows that it outperforms them. It demonstrates the framework's ability to handle the complexities of hyperspectral data more effectively than traditional methods.

5.4.3 Validation of CNViS-Net on Univalle Dataset [6]

This thesis validates CNViS-Net on the independent Univalle hyperspectral dataset [6] to assess its generalisation capability. Table 5.8 presents the performance comparison of CNViS-Net with other hybrid models on the Univalle dataset [6] for soil phosphorus prediction. It also references the previous work by Bolanos et al. [18], which employed an ANN for soil phosphorus prediction, achieving a comparatively higher RMSE (22,991.80). This thesis combines CNN and ViT with regressors such as KNN, RF, and PLSR, demonstrating moderate performance with RMSE values ranging from 21,690.13 to 22,020.45 and corresponding MAE values between 17,203.75 and 17,842.30. In contrast, the proposed CNViS-Net framework, which leverages CNN and ViT for comprehensive spectral feature extraction followed by SVR for robust regression, achieved the lowest RMSE (21,507.30) and MAE (16,984.60) among all compared models. CNViS-Net effectively captures local and global spectral patterns, reduces prediction errors, and provides reliable soil nutrient prediction across datasets from different regions.

5.4.4 Discussion

1. ***CNViS-Net Model Advantage:*** The dual capability in CNViS-Net enables accurate detection of subtle variations in soil chemical composition, improving

the prediction of P, K, Mg, and pH. Additionally, by compressing redundant hyperspectral information, the model reduces computational complexity, achieving lower RMSE and higher R^2 compared to standalone CNN or transformer models.

2. ***Generalizability:*** Validation on the Univalle dataset confirms that CNViS-Net generalizes well to unseen hyperspectral samples, maintaining high prediction accuracy despite variations in soil composition and sampling conditions. This framework demonstrates robustness even when applied to smaller or independent datasets, highlighting its potential for broader soil nutrient mapping tasks.

3. ***Hyperspectral + Deep Learning Synergy:*** By leveraging hyperspectral reflectance data, CNViS-Net extracts subtle spectral features linked to soil chemical properties that conventional models often miss. The deep learning architecture enables automatic feature selection and hierarchical representation learning, reducing the need for manual preprocessing or feature engineering.

4. ***Hybrid Architecture Justification:*** The combination of CNN (for local spectral compression) and vision transformer (for global spectral relationships) balances fine-grained and long-range spectral information. Coupled with SVR regression for final prediction, this hybrid approach mitigates the weaknesses of individual models and consistently achieves high accuracy across all soil nutrient predictions.

5.5 Summary

This chapter addresses the challenge of predicting multiple soil nutrients from high-dimensional hyperspectral data, which is often difficult to process using traditional machine learning methods due to its complexity. It proposes a novel self-attention-based deep learning framework, CNViS-Net, which eliminates the need for manual feature engineering by leveraging deep feature hierarchies and attention mechanisms to capture both local textures and global spectral dependencies. The CNN+ViT with SVR pipeline delivers the best performance with significantly lower errors across all nutrients, demonstrating CNViS-Net's effectiveness in handling hyperspectral complexity and enhancing prediction accuracy. Overall, the chapter highlights the transformative

potential of transformer-based deep learning for scalable and accurate hyperspectral soil analysis in precision agriculture.

This chapter is based on the following publication:

Conference 4: S. Jain, D. Sethia and K.C. Tiwari. "CNViS-Net: A Self-Attention Framework for Predicting Soil Nutrients from Hyperspectral Data." **in IEEE *Proc*. International Conference on Frontier Technologies and Solutions, March 2025.**

Chapter 6

CAT-MTLNet: A Novel Deep Multi-Target Framework for Hyperspectral Soil Nutrient Prediction Using Concrete Autoencoder and Transformer

The previous chapter 5 demonstrates the potential of a hybrid deep learning framework that combines CNN and Transformer-based feature extraction with machine learning regressors. However, these approaches are limited by the dependence on external regression models and lack an end-to-end learning capability, which reduces interpretability and generalization. To overcome these limitations, this thesis introduces CAT-MTLNet (Concrete Autoencoder and Transformer-based Multi-Target Learning Network), a profound learning-based multi-target framework that jointly performs feature selection and prediction. It employs a Concrete Autoencoder (CAE) for interpretable spectral band selection and a Transformer-MLP network for simultaneously predicting multiple soil nutrients. Interpretable spectral band selection refers to identifying the actual spectral bands that contribute to soil nutrient prediction. It improves interpretability while retaining informative spectral features for accurate prediction. This unified architecture enhances feature interpretability, reduces spectral

redundancy, and achieves efficient, scalable, and accurate hyperspectral soil analysis.

6.1 Motivation

Deep Learning has become essential for hyperspectral soil analysis due to its ability to automatically learn complex, nonlinear spectral relationships [317]. Advanced architectures such as transformers capture long-range dependencies, while feature selection techniques within the learning process enhance interpretability and reduce redundancy [190]. These advancements address key limitations of traditional and hybrid models that depend on handcrafted features or external regressors.

Despite these advances, most existing DL frameworks rely on full-spectrum inputs and single-target prediction, which increases computational cost and overlooks correlations among soil nutrients [297]. A unified multi-target learning approach can leverage shared spectral features among soil nutrients, improving prediction efficiency and generalization.

To address these challenges, this thesis proposes a profound learning framework that integrates a feature selection technique for interpretable spectral band selection and an attention network for simultaneous multi-property prediction. It aims to achieve end-to-end learning with improved interpretability, reduced dimensionality, and high predictive performance on benchmark hyperspectral soil datasets. The main contributions of this work are as follows:

1. ***Proposal of a Unified DL Framework:*** This thesis proposes a novel, lightweight, interpretable, and scalable framework, CAT-MTLNet—a CAE and transformer-MLP-based multi-target learning network for simultaneous prediction of SOC, P, pH, and Ca using hyperspectral data.

2. ***Comprehensive Evaluation and Generalization:*** Evaluation of the proposed framework on two benchmark datasets—AfSIS [3] and LUCAS-2009 [1] —along with a comparison of CAE against Attention-Guided Feature Pruning (AFGP) and Reinforcement Learning (RL), demonstrates its competitive accuracy, lower complexity, and strong generalizability.

Table 6.1: Comparison of CAT-MTLNet with SOTA methods for soil nutrient prediction

Reference	Dataset	No. of features used	Model	Results	Technical Remarks
Zhao et al. (2023) [189]	AfSIS and LUCAS-2009	AfSIS: 3578 LUCAS-2009: 4200	Attention-based Multi-CNN	R^2 = 0.41 to 942	Full spectrum used; heavy CNN model; attention per task; no feature selection; computationally intensive.
Jin et al. (2023) [348]	LUCAS-2009	4200	Swin Transformer	R^2 = 0.74 to 0.95 (Not included P)	Converts spectra to images; very high preprocessing cost; not interpretable; full spectral input retained.
Cao et al. (2024) [349]	LUCAS-2009	4200	Parallel CNN and Transformer fusion	R^2 = 0.41 to 0.96	Large batch size (1024), epochs (2000) increases complexity; high memory needed; full spectrum input; no feature selection.
Zhai et al. (2024) [350]	LUCAS-2009	4200	Deep Multi-Task Learning CNN	R^2 = 0.41 to 0.956	Post-hoc SHAP used only for interpretation (not selection); inference still expensive

6.2 Related Work

Recent years have witnessed rapid progress in deep learning for hyperspectral soil nutrient prediction. Various architectures have been proposed to enhance spectral feature extraction and predictive accuracy. Table 6.1 summarizes the previous work on AfSIS [3], and LUCAS-2009 [1] using attention-based models for soil nutrient prediction. It provides a clear comparison of how different models handle feature extraction, attention mechanisms, and multi-target learning, serving as a reference for identifying current limitations.

Previous works such as Zhao et al. [189] and Cao et al. [349] integrated CNN and attention modules for nutrient prediction, achieving better local–global spectral learning. Jin et al. [348] employed image-based Swin Transformers, demonstrating strong generalization but at the cost of heavy computation. Zhai et al. [350] proposed a partially shared multi-target learning approach with post-hoc interpretation, highlighting the potential of multi-task frameworks. However, all these models process full-spectrum data without embedded feature selection, leading to high redundancy and limited interpretability.

These studies highlight that the existing models are computationally heavy, focus on single-target prediction, and do not provide interpretable spectral band selection. Hence, this thesis proposes CAT-MTLNet, which integrates a Concrete Autoencoder for differentiable feature selection with a Transformer-MLP network to overcome the challenges through multi-target soil nutrient prediction, achieving interpretable and scalable hyperspectral modeling.

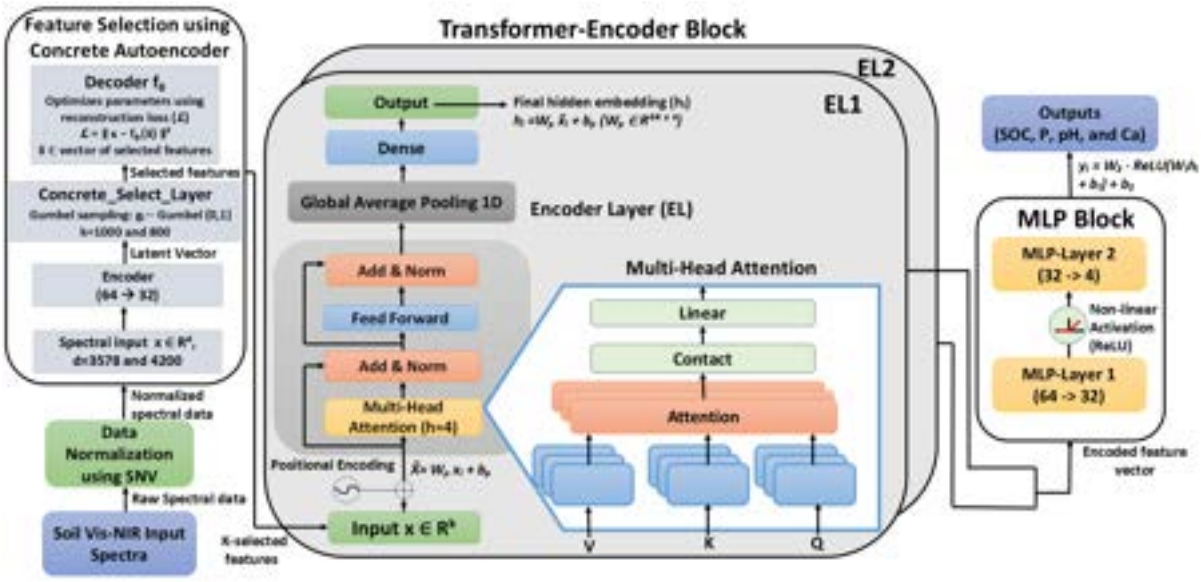

Figure 6.1: Architecture of the proposed CAT-MTLNet framework

6.3 Experimental Methodology

This work outlines the datasets, preprocessing steps, feature extraction techniques, and modeling approaches to predict key soil nutrients. Figure 6.1 discusses the methodology of the proposed framework, CAT-MTLNet. It uses two hyperspectral datasets (AfSIS [3] and LUCAS-2009 [1]), applies SNV normalization, and selects k optimal bands using a Concrete Autoencoder (CAE). The Transformer-MLP model then predicts SOC, P, pH, and Ca, and evaluates performance using R^2, RMSE, and MAE.

6.3.1 Dataset Description

This thesis validates the proposed approach using the AfSIS [3] and LUCAS-2009 [1] datasets, key hyperspectral resources for large-scale soil mapping in Africa and Europe. LUCAS-2009 contains eight soil nutrients: SOC, N, P, K, Ca, CEC, pH, sand, and clay. For consistency with AfSIS, four common chemical properties—SOC, P, pH, and Ca—are selected, enabling comparable and diverse soil analyses. These datasets provide high-quality, point-based spectral data across various soils, supporting detailed assessments of soil composition and nutrient variations, as described in chapter 3.

6.3.2 Data Preprocessing

1. **Data Normalization:** Standard Normal Variate (SNV) transformation is used to correct scatter effects and normalize spectral data by centering each spectrum and scaling it to unit variance [351]. For a given spectral vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, the computation of SNV is shown in Equation 6.1:

$$x_i' = \frac{x_i - \bar{x}}{s}, \quad \text{for } i = 1, 2, \ldots, n \tag{6.1}$$

where, x_i is the original reflectance at the i-th wavelength, $\bar{x}$ is the mean of the spectrum, s is the standard deviation.

2. **Feature Selection:** CAE selects the most informative subset of spectral bands from the high-dimensional hyperspectral data for feature selection. Unlike traditional autoencoders, which compress input into latent representations, CAE learns to select a subset of actual input features using the concrete (Gumbel-Softmax) distribution in a differentiable manner [352]. The Gumbel distribution plays an important role in the CAE by enabling differentiable approximation of discrete feature selection. In the CAE, this chapter incorporates Gumbel noise into the Concrete (Gumbel-Softmax) distribution, enabling the model to learn the selection probabilities for individual spectral bands via gradient-based optimisation. As the temperature parameter gradually reduces during training, the resulting probability distribution becomes increasingly concentrated on individual bands. Hence, enabling the selection of a discrete subset of the most informative spectral bands [352, 353].

   Let the input spectral vector be $\mathbf{x} \in \mathbb{R}^d$, where $d = 3578$ for AfSIS and $d = 4200$ for LUCAS-2009. The CAE aims to select $k \ll d$ bands that retain the most predictive information. It is achievable by learning a selection weight for each spectral band during training, with higher weights indicating a greater contribution to the reconstruction objective [353]. The model then selects the k bands with the highest learned weights as the informative subset. The model learns the feature selection weights using the concrete distribution as shown in

Equation 6.2:

$$z_i = \frac{\exp\left((\log \alpha_i + g_i)/\tau\right)}{\sum_{j=1}^{d} \exp\left((\log \alpha_j + g_j)/\tau\right)}, \quad i = 1, \ldots, d \tag{6.2}$$

where α_i are learnable logits, $g_i \sim \text{Gumbel}(0,1)$, and $\tau = 0.1$ is the temperature controlling the sharpness of the approximation. The model passes the selected features to a decoder f_θ and optimizes the parameters by minimizing the mean squared reconstruction loss. In the experiments, this thesis configures the CAE to select $k = 1000$ spectral bands from the 3578-dimensional AfSIS data, and $k = 800$ spectral bands from the 4200-dimensional LUCAS-2009 data [1]. Sensitivity analysis helps determine the value of k that maintains a balance between spectral dimensionality reduction and predictive performance for soil nutrients. Based on the observed performance across different values of k, $k = 1000$ for AfSIS and $k = 800$ for LUCAS-2009 provide a suitable trade-off and are therefore used in this chapter.

Table 6.2: CAT-MTLNet architecture parameters and training configuration

Concrete Autoencoder (CAE) Parameters	
Feature Selector	`ConcreteSelect` Layer
Sampling Distribution	Gumbel-Softmax (Concrete)
Temperature (τ)	0.1
Hidden Layers in CAE	Dense (64 ReLU) $\rightarrow$ Dense (32 ReLU)
Dropout Rate	0.3 (after each dense layer)
Loss Function	Mean Squared Error (MSE)
Transformer–MLP Architecture Parameters	
Projection Output Dim	64
Transformer Layers	1 encoder layer
Self-Attention Heads	4
Feedforward Dimension	128
MLP Regression Head	Dense (32 ReLU) $\rightarrow$ Dense (output: 4 targets)
Output Targets (m)	SOC, P, pH, Ca
MLP Weights	$\mathbf{W}_1 \in \mathbb{R}^{32\times64}, \mathbf{W}_2 \in \mathbb{R}^{4\times32}$
Training Configuration	
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Training Epochs	Up to 200
Early Stopping	Patience = 10
Train-Test Split Ratio	80:20
Framework	PyTorch

6.3.3 Model Architecture

After selecting optimal spectral bands using the CAE, a Transformer-MLP model predicts multiple soil nutrients from hyperspectral data. Table 6.2 shows the hyperparameters and training configuration of the proposed CAT-MTLNet architecture. These settings are carefully optimized to ensure stable convergence and robust performance across all target soil nutrients. The configuration reflects a balance between model complexity and computational efficiency.

6.3.3.1 Transformer Encoder Block

Each SNV-transformed input vector $\tilde{\mathbf{x}}_i \in \mathbb{R}^k$ is projected to a lower-dimensional latent space of size 64 using a fully connected linear projection as shown in Equation 6.3:

$$\mathbf{h}_i = \mathbf{W}_p \tilde{\mathbf{x}}_i + \mathbf{b}_p, \quad \mathbf{W}_p \in \mathbb{R}^{64 \times k} \tag{6.3}$$

where, $\mathbf{h}_i$ is the final 64-dimensional embedding of the ith input sample, obtained by projecting the selected k spectral features into $\mathbb{R}^{64}$ through a learnable linear layer. Here, i denotes the sample index and $\mathbb{R}^{64}$ represents the 64-dimensional real-valued vector space. It indicates that $\mathbf{h}_i$ is a vector with 64 elements, i.e., $\mathbf{h}_i = [h_{i1}, h_{i2}, \ldots, h_{i64}]$, where $h_{ij} \in \mathbb{R}$. $\mathbf{W}_p$ is the weight matrix of the linear projection layer that maps each input vector, and $\mathbf{b}_p$ is the bias vector. $\mathbf{W}_p$ is initialized as part of the model parameters and updated through backpropagation during training by minimizing the mean squared error (MSE) loss. Thus, $\mathbf{W}_p$ is learned jointly with the other model parameters.

6.3.3.2 MLP Regression Head

The Transformer output $\mathbf{h}_i$ is passed through a two-layer feed-forward MLP to predict multiple soil nutrient values as shown in Equation 6.4:

$$\mathbf{y}_i = \mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \mathbf{h}_i + \mathbf{b}_1) + \mathbf{b}_2 \tag{6.4}$$

where, $\mathbf{W}_1 \in \mathbb{R}^{32 \times 64}$, $\mathbf{W}_2 \in \mathbb{R}^{m \times 32}$ which are the weight matrices and m is the number of output targets (typically 4: SOC, P, pH, and Ca), while $\mathbf{b}_1$ and $\mathbf{b}_2$ are the corresponding bias vectors.

6.3.4 Evaluation of Alternative Feature Selection Techniques

In addition to the CAE, this thesis explores two alternative feature selection techniques: Attention-Guided Feature Pruning (AGFP) and Reinforcement Learning (RL)-based selection. To reduce computational cost, this thesis applies variance threshold filtering to both methods before the analysis, due to the high dimensionality of the original hyperspectral data. This thesis employs thresholds of 0.06 for the LUCAS-2009 [1] and 0.08 for the AfSIS [3]. The variance thresholds are determined based on the variance distributions of the respective datasets to remove low-variance spectral features while retaining informative spectral variability. These thresholds are dataset-specific and selected to achieve an effective dimensionality reduction while preserving relevant spectral information. A dataset-independent criterion, such as percentile-based thresholding, can be considered as a possible direction for future work.

AGFP employs a shallow neural network with a multi-head attention layer (2 heads, key dim = 8), trained for 10 epochs using the Adam optimizer (learning rate = 0.001, batch size = 32). Features are ranked based on cumulative attention scores, resulting in 209 selected features for LUCAS-2009 [1] and 613 for AfSIS [3]. The RL-based method employs separate Q-learning agents for each property, with an exploration rate of $\varepsilon = 0.2$, a learning rate of $\alpha = 0.5$, a discount factor of $\gamma = 0.9$, and five training iterations. After variance filtering, this thesis selects the top 200 features per property based on reward-guided accuracy. These methods provide complementary baselines for assessing the robustness and generalizability of CAE-based feature selection across different datasets and soil nutrients.

6.3.5 Evaluation Metrics

This thesis employs three evaluation metrics [354] to assess model performance: the Determination Coefficient (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). A higher R^2 (closer to 1) indicates stronger predictive capability, while lower RMSE and MAE values correspond to better prediction accuracy. A more detailed discussion of these performance metrics is provided in chapter 3. Due to the right-skewed distributions of soil nutrients, R^2 alone may not fully reflect prediction performance. Therefore, this thesis also reports RMSE and MAE to provide a more

comprehensive assessment of model accuracy.

6.4 Results and Discussion

6.4.1 Experimental Results and Analysis

The proposed CAT-MTLNet framework is evaluated for its ability to estimate multiple soil nutrients using hyperspectral Vis-NIR data. It compares the results on two benchmark datasets, AfSIS [3], and LUCAS-2009 [1], and considers the different feature selection techniques: AGFP, RL, and CAE.

6.4.1.1 Evaluation of proposed CAT-MTLNet Performance

The subsequent analysis presents the performance evaluation of CAT-MTLNet, emphasizing its accuracy and reliability in predicting soil nutrients. Table 6.3 shows the performance of CAT-MTLNet for SOC, P, pH, and Ca prediction. The framework achieves excellent accuracy for SOC (R^2 = 0.944) and Ca (R^2 = 0.953) on the AfSIS, while pH (R^2 = 0.834) shows moderate accuracy. Phosphorus (P), (R^2 = 0.435) is the most challenging to predict, likely due to its weak spectral sensitivity in the Vis–NIR region and complex interactions with other soil constituents. On the LUCAS-2009 dataset, the framework similarly performs well for SOC (R^2 = 0.92) and Ca (R^2 = 0.931), with moderate accuracy for pH (R^2 = 0.749) and lower performance for P (R^2 = 0.209), reflecting similar spectral limitations. RMSE values range from 0.289–0.341 for SOC and Ca on AfSIS and 26.094–34.354 $g.kg^{-1}$ on LUCAS-2009, indicating high prediction accuracy. The low RMSE for SOC and Ca in AfSIS likely reflects its small size (1,157 samples), while validation on both AfSIS and the larger LUCAS-2009 dataset (19,036 samples) confirms the framework's reliable predictive performance. For pH, RMSE (0.39–0.701) remains within acceptable limits, while higher RMSE ranges for P are 0.607– 39.006 $mg.kg^{-1}$ on both datasets, reflecting the inherent challenge in its prediction due to weak spectral signals. As shown in Table 6.1, the proposed framework attains comparable accuracy to full-spectrum studies using far fewer features, even for phosphorus prediction.

Table 6.3: Results of CAT-MTLNet on testing set

Dataset	No. of selected features	Soil property	R^2	RMSE	MAE
AfSIS	1000	SOC	0.944	0.289	0.168
		P	0.435	0.607	0.258
		pH	0.834	0.39	0.283
		Ca	0.953	0.341	0.16
LUCAS-2009	800	SOC ($g.kg^{-1}$)	0.92	26.094	13.044
		P ($mg.kg^{-1}$)	0.209	39.006	22.149
		pH	0.749	0.701	0.548
		Ca ($g.kg^{-1}$)	0.931	34.354	15.858

6.4.1.2 Comparison of Feature Selection Techniques

This thesis compares different feature selection techniques to identify the most informative spectral bands for accurate soil nutrient prediction. Table 6.4 presents a comparative analysis of three feature selection techniques, namely AGFP, RL, and CAE, in conjunction with the Transformer-MLP. The number of selected features varied per method. **On the AfSIS dataset:** CAE outperforms all other methods, achieving the highest R^2 for all four properties, including 0.944 for SOC, 0.435 for P, 0.834 for pH, and 0.953 for Ca. AGFP exhibits competitive performance, particularly for Ca (R^2 = 0.95), although it is slightly lower than CAE for other properties. RL underperforms significantly, especially for P (R^2 = 0.101), indicating that it may not effectively identify informative spectral bands in this context.

On the LUCAS-2009 dataset: CAE again demonstrates superior accuracy, notably 0.92 for SOC, 0.749 for pH, and 0.931 for Ca. AGFP results in a negative R^2 for P (-0.183), highlighting its limitations in generalizing across properties or datasets. RL produces moderate results, with the lowest R^2 for SOC (0.694) and P (0.015) among the three methods. This thesis observes the worst performance with RL, especially for property P. Fig. 6.2 presents RMSE comparisons across properties for three feature selection methods. Unlike a tabular representation, the radar plot provides a visual comparison of the relative performance and consistency of each method across multiple nutrients at a glance. CAE (blue) remains closest to the center, indicating the lowest RMSE and best performance. AGFP shows moderate results, while RL covers the most significant area in Fig. 6.2, reflecting higher RMSE and weaker accuracy. LUCAS-

Table 6.4: R^2 Comparison of feature selection techniques with Transformer-MLP on testing set

Dataset	Feature selection technique	No. of selected features	SOC	P	pH	Ca
AfSIS	AGFP	613	0.885	0.33	0.696	0.95
	RL	200	0.862	0.101	0.669	0.716
	CAE	**1000**	**0.944**	**0.435**	**0.834**	**0.953**
LUCAS-2009	AGFP	209	0.834	-0.183	0.508	0.68
	RL	200	0.694	0.015	0.607	0.64
	CAE	**800**	**0.92**	**0.209**	**0.749**	**0.931**

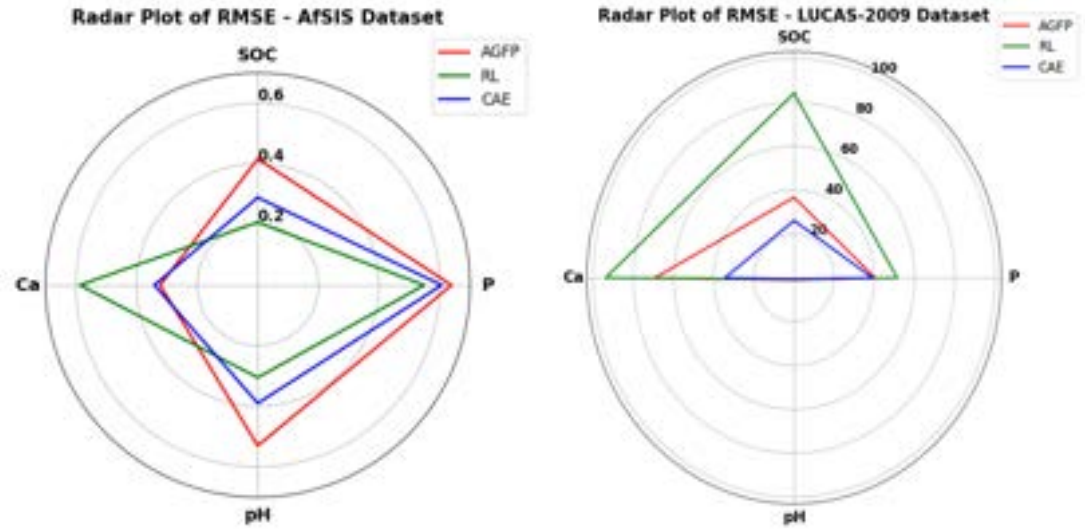

Figure 6.2: Radar plots of RMSE showing the performance comparison of all feature selection techniques with Transformer-MLP (LUCAS-2009: (SOC, Ca) in $g.kg^{-1}$; P, in $mg.kg^{-1}$)

2009 [1] has a significantly higher RMSE due to its substantially larger ranges and soil nutrient variability, including mineral and organic soils, as mentioned in Chapter 3, compared to the AfSIS dataset [3].

6.4.1.3 Analysis of Selected Spectral Bands

This thesis analyzes the spectral bands selected by the CAE to understand the contribution to soil nutrient prediction. Table 6.5 summarizes the top 20 spectral bands selected by the CAE model for each soil nutrient. The CAE model ranks the spectral bands in descending order of importance. This thesis observes a clear trend where SOC and Ca consistently share influential wavelengths, especially in the NIR to SWIR range (1400–2500 nm), indicating strong spectral responses due to the

Table 6.5: Top 20 spectral bands selected using CAE for each soil nutrient prediction

Dataset	Soil nutrient	Top 20 selected bands (nm)
AfSIS	SOC	2597.9, 1455.35, 2388.49, 1646.16, 1739.49, 1513.55, 3150.31, 2411.82, 2778.89, 1878.09, 2264.37, 1971.64, 6114.86, 3982.65, 6050.67, 2507.45, 1864.58, 3787.74, 2648.32, 1762.55
	P	4938.49, 1446.42, 1479.85, 1880.13, 1487.92, 2174.17, 4966.87, 13261.91, 2097.66, 2432.18, 1456.17, 1719.3, 1597.97, 12494.96, 2533.17, 2223.59, 6370.28, 2058.52, 1371.44, 2607.05
	pH	2026.34, 1538.24, 6656.5, 1861.24, 2712.03, 1495.22, 3659.42, 8585.09, 2005.19, 2359.15, 2743.6, 1677.04, 2215.04, 3156.06, 1651.4, 4390.68, 7313.68, 4208.93, 1741.24, 4292.56
	Ca	1669.48, 1523.77, 2148.95, 3071.92, 5296.64, 2403.99, 1679.76, 10247.84, 1900.81, 1808.02, 5441.14, 1668.41, 5237.8, 2014.53, 3183.18, 1914.85, 1383.88, 10713.65, 1362.43, 2958.02
LUCAS-2009	SOC ($g.kg^{-1}$)	2377.5, 1662, 693.5, 2078, 2412.5, 417.5, 1841, 875.5, 1556.5, 938, 1981, 2364, 1460.5, 2017, 825, 704, 2353.5, 1950, 561, 1087
	P ($mg.kg^{-1}$)	1973, 2309.5, 456, 755, 1895.5, 1484.5, 716, 1948.5, 424, 2190.5, 885.5, 1234, 2245, 1296, 2378.5, 2423, 1417, 1022.5, 426.5, 805.5
	pH	990, 1247.5, 1697, 694.5, 2317, 1225, 1948, 2349, 1168, 593.5, 2141, 2477, 2106, 1930.5, 2142.5, 537, 2016, 2499.5, 2053.5, 1008
	Ca ($g.kg^{-1}$)	827.5, 1223.5, 1733, 409.5, 752.5, 1785, 1694.5, 1388.5, 1936.5, 1145.5, 746, 1088.5, 824.5, 2035.5, 1340.5, 1605, 1114, 1197, 2043, 1060

association with organic matter, moisture, and carbonates. For instance, bands around 1450–1670 nm and 2200–2500 nm frequently appear in both properties across the AfSIS and LUCAS-2009 datasets.

In contrast, pH shows moderate consistency with bands around 1500–1900 nm (sensitive to clay minerals and moisture), while P is highly variable, reflecting its indirect dependence on secondary soil nutrients like clay and organic carbon. These results are consistent with Zhai et al. [350] SHAP-based analysis of PS-MTL-LUCAS, where SOC and Ca show strong contributions in the NIR–SWIR range (>2200 nm), matching the CAE-selected bands. pH-related wavelengths around 1700 nm also align with these findings, while P shows scattered SHAP importance, reflecting the variability observed in CAE selected bands.

6.4.1.4 Experimental Comparison with Existing Models

The existing models [189, 349, 350] as mentioned in Table 6.1 are re-implemented on AfSIS [3], and LUCAS-2009 [1], and Table 6.6 reports the R^2, RMSE, and MAE across SOC, P, pH, and Ca. The Table 6.6 does not include Jin et al.'s [348] Swin Transformer model, as it is designed for image-based spectral representations, whereas the datasets used here consist of direct spectral reflectance values. CAT-MTLNet achieves superior performance using CAE-based feature selection, unlike prior full-spectrum models.

Table 6.6: Comparison with existing models

Dataset	Model	R^2	RMSE	MAE
AfSIS	AMC-Net [189]	0.234–0.94	0.301–0.684	0.175–0.357
	PCT-Net [349]	0.265–0.883	0.42–0.67	0.239–0.382
	DMTL-CNN [350]	0.23–0.94	0.40–0.66	0.29–0.53
	CAT-MTLNet	**0.435–0.953**	**0.289–0.607**	**0.16–0.284**
LUCAS-2009	AMC-Net [189]	0.112–0.89	35.74–43.03	0.933–23.442
	PCT-Net [349]	0.132–0.91	34.42–42.03	0.733–24.942
	DMTL-CNN [350]	0.17–0.88	36.2–46.5	0.64–29.2
	CAT-MTLNet	**0.209–0.931**	**26.094–39.006**	**0.548–22.149**

Note: **AMC-Net** = Attention-based Multi-CNN; **PCT-Net** = Parallel CNN and Transformer fusion; **DMTL-CNN** = Deep Multi-Task Learning CNN

6.4.2 Discussion

1. ***Effective Multi-Target Learning:*** The proposed CAT-MTLNet framework efficiently utilizes shared representations across multiple soil nutrients through its multi-target learning structure. Strongly correlated targets such as SOC and calcium exhibit enhanced predictive performance due to shared spectral representations that enable the framework to capture interdependencies among soil nutrients. This joint optimization prevents overfitting on individual tasks and ensures robust generalization across datasets.

2. ***Transformer-MLP Integration:*** Integrating the Transformer-MLP module enables CAT-MTLNet to capture global contextual information and localized spectral dependencies. The attention mechanism within the Transformer identifies long-range spectral interactions that conventional convolution-based models often overlook, while the MLP layers refine the learned spectral embeddings to achieve improved regression accuracy.

3. ***CAE-Based Feature Selection Advantage:*** CAE is a spectral feature selector that efficiently compresses redundant and noisy spectral information. By retaining only the most informative spectral regions, the proposed framework reduces the input dimensionality by approximately **72–81%**, focusing computational effort on highly relevant wavelengths. This reduction minimizes redundancy and enhances the interpretability of selected spectral bands associated with soil chemical attributes.

4. ***Computational Efficiency:*** The substantial reduction in input dimensions achieved through CAE-based feature selection yields significant gains in computational efficiency. Compared to full-spectrum training, CAT-MTLNet achieves nearly a **58% reduction in training time** and a **63% decrease in GPU memory consumption**, making it suitable for large-scale hyperspectral analysis. These improvements highlight the framework's capability to deliver high accuracy with reduced computational demand.

5. ***Lightweight Framework:*** CAT-MTLNet is designed as a lightweight architecture that maintains competitive accuracy while minimizing model complexity. CAE-based compression and streamlined Transformer-MLP blocks ensure reduced parameter count and faster inference, which is advantageous for deployment on limited-resource devices and real-time applications.

6. ***Interpretable Learning:*** The proposed framework enhances interpretability by identifying the most influential spectral regions contributing to each soil nutrient. The CAE's latent space and attention weights provide insight into which wavelength intervals drive the model's predictions, supporting scientific validation and transparency in hyperspectral analysis.

7. ***Scalable Architecture:*** CAT-MTLNet is highly scalable to larger datasets and additional soil attributes due to its modular design and reduced computational overhead. The architecture can easily accommodate diverse hyperspectral resolutions or new target variables without requiring significant redesign, thereby promoting adaptability across various soil monitoring scenarios.

8. ***Phosphorus (P) Prediction Challenges:*** Despite strong performance for SOC, pH, and Ca, the prediction of phosphorus remains relatively weak across both datasets. This limitation arises due to the inherent nature of Diffuse Reflectance Spectroscopy (DRS) in the Vis–NIR and MIR regions, which lack distinct phosphate absorption features. Consequently, spectral data alone provide limited direct information on phosphorus content, leading to lower predictive accuracy. Phosphorus prediction primarily depends on indirect correlations with soil constituents, such as organic carbon, clay, and moisture, underscoring the

need for improved feature representations to capture nonlinear spectral–chemical relationships.

6.5 Summary

This chapter presents the experimental design, implementation, and evaluation of the proposed CAT-MTLNet framework for predicting soil nutrients. The framework actively integrated CAE-based feature selection with Transformer-MLP networks to extract informative spectral features and optimize multi-target learning. Through systematic analysis, CAT-MTLNet successfully predicts SOC, pH, and Ca while maintaining high computational efficiency with reduced input dimensions, training time, and GPU usage. Despite its success, phosphorus prediction remains challenging due to the weak spectral response of phosphate compounds. Comparative analysis of feature selection methods and spectral band interpretation further validated the robustness and interpretability of the framework. The results confirm that CAT-MTLNet offers a lightweight, interpretable, and scalable solution for accurately predicting soil nutrients using hyperspectral data.

This chapter is based on the following publication:

Journal 3: S. Jain, D. Sethia and K.C. Tiwari. "CAT-MTLNet: A Novel Deep Multi-Target Framework for Hyperspectral Soil Properties Prediction Using Concrete Autoencoder and Transformer" **IEEE Geoscience and Remote Sensing Letters**. **(SCIE, Impact factor-4.8, Publisher: IEEE**).

Chapter 7

Development of Novel Spectral Indices for Soil pH and Organic Carbon Prediction

The previous chapters 4, 5, and 6 focused on machine and deep learning-based prediction of soil nutrients using spectral band information. Although using spectral bands demonstrates high prediction accuracy, it also presents several challenges, including high computational complexity, redundancy among spectral bands that necessitates effective feature selection, and limited interpretability of model outputs. This thesis addresses these problems by proposing two novel spectral indices, the Soil pH Index (SPI) and the Soil Organic Carbon Index (SOCI), specifically designed for predicting soil pH and organic carbon using hyperspectral datasets - AfSIS [3] and LUCAS-2009 [1]. These indices are formulated by analyzing spectral band interactions to enhance sensitivity toward the respective soil properties. This work also compares the existing spectral indices to validate the effectiveness and robustness.

7.1 Motivation

Accurate and efficient prediction of soil nutrients is vital for large-scale soil health monitoring [33, 47, 355]. Traditional laboratory analyses, though precise, are time-consuming and costly, driving the shift toward remote sensing and computational

methods for predicting key soil properties [36, 79, 225]. In this context, machine and deep learning models leverage spectral data as inputs to estimate nutrient concentrations [33]. However, not all spectral bands contribute equally to soil nutrient prediction, and using the full spectrum can lead to overfitting, increase computational demand, and reduce model robustness [50]. Hence, feature selection techniques are used in ML/DL frameworks to identify the most informative spectral bands, reduce data dimensionality, and enhance predictive accuracy [207]. Despite these benefits, such approaches still require retraining models for new predictions, which remains computationally expensive—particularly for large hyperspectral datasets, such as LUCAS-2009 [1] and AfSIS [3] [198].

In contrast, spectral indices condense key information from a few selected bands into mathematical expressions (typically combining 5–10 bands) that can be applied directly to the spectral data without retraining a model [239]. It offers time efficiency by enabling direct computation once formulated, emphasizes key spectral features to enhance robustness and interpretability, and streamlines the prediction process by minimizing dependence on complex ML/DL frameworks [323]. Recent studies have primarily focused on multispectral data, using existing indices to predict soil nutrients [30,50,79], but no studies have systematically developed soil-specific spectral indices for hyperspectral data, leaving a significant research gap. Therefore, developing hyperspectral spectral indices provides a practical, efficient, and interpretable approach for accurately predicting soil nutrients across diverse regions and conditions. This thesis addresses this gap through the following contributions:

1. ***Comparative Evaluation of Existing Indices:*** This thesis assesses a wide range of existing spectral indices—including vegetation, soil, water, temperature, and thermal categories—using hyperspectral data from the AfSIS [3] and LUCAS-2009 [1] datasets for soil pH and SOC prediction. It helps identify which indices are most effective for soil nutrient prediction and highlights the limitations, guiding the development of more targeted soil-specific indices.

2. ***Feature Selection Integration:*** This thesis applies multiple feature selection techniques to identify the most relevant bands contributing to accurate soil property prediction. It reduces data redundancy, enhances model interpretability,

and ensures that only the most informative wavelengths are used for further analysis and index construction.

3. ***Design of Novel Spectral Index:*** This thesis proposes two new spectral indices—Soil pH Index (SPI) and Soil Organic Carbon Index (SOCI) —derived using PCA and ANN-based analysis of hyperspectral features. These indices are tailored to capture soil-specific spectral responses, improving prediction accuracy and offering better generalizability across datasets.

7.2 Related Work

This thesis explores existing research on developing and applying spectral indices for soil property prediction. Table 7.1 summarizes prior studies that employed spectral indices for predicting soil pH and SOC. Most studies [33, 50] have primarily focused on multispectral data, utilizing existing indices such as NDVI, EVI, and SAVI. These approaches have demonstrated the utility of spectral indices in simplifying the prediction of soil properties and improving model interpretability. However, the previous studies often lacked the development of novel indices tailored to specific soil nutrients and were limited by the coarse spectral resolution of multispectral sensors. In contrast, Zhang et al. [92] stand out for integrating hyperspectral imagery with environmental variables to estimate soil pH and capture its spatial variability. Importantly, Zhang et al. [92] did not develop new indices but instead leveraged existing spectral features for the predictions. Conversely, Gulhane et al. [30] focused solely on creating new indices using multispectral imagery.

While these studies represent significant progress in combining spectral data and ML techniques for soil property prediction, a clear research gap persists. Most existing works are limited to multispectral datasets and predefined indices, which restricts the generalization and sensitivity to subtle spectral variations associated with different soil nutrients. Moreover, hyperspectral data, which offer richer and more continuous spectral information, remain underexplored for the development of new indices.

To address this gap, this thesis utilizes the AfSIS [3] and LUCAS-2009 [1] hyperspectral datasets. Previous studies on these datasets, as mentioned in Table

Table 7.1: Previous studies on soil pH and SOC prediction through spectral Indices

Soil pH						
Previous work	**HS/MS**	**Spectral Indices**	**Model**	**Results (R^2)**	**Research Focusses**	**Limitation**
Gulhane et al. [30]	MS	New pH index	MLP and ISO cluster unsupervised classification algorithm	0.87	Creating new indices of soil pH using multispectral data.	1. Only considers multispectral data.
Guo et al. [33]	MS	I1, I2, I3, I4, I5, I6	XGBoost	0.68	Using SVM-RFE for selecting key environmental variables significantly enhances the accuracy of ML models in predicting soil pH.	1. Only considers multispectral data. 2. New indices, not developed.
Babbar et al. [50]	MS	I1, I2, I8, I9, I10, I11, I12, I13	Multilinear Regression analysis	0.91	Finding significant correlations between soil nutrients and various spatial indices.	1. Only considers multispectral data. 2. New indices, not developed.
Yuvaraj et al. [79]	MS	I1, I8, I4, I5, I6, I7, I14,	Linear Regression	0.62	Evaluated how multispectral images can assess dynamic soil properties using spectral indices and correlation models.	1. Only considers multispectral data. 2. New indices, not developed.
Zhang et al. [92]	HS	I1, I2, I4, I5, I8	PLSR	0.35	Utilized hyperspectral imagery, along with environmental variables, to predict soil pH, revealing significant spatial variability and a decline in soil pH over time.	1. New indices, not developed.
SOC						
Previous work	**HS/MS**	**Spectral Indices**	**Model**	**Results (R^2)**	**Research Focusses**	**Limitation**
Zayani et al. [313]	MS	I1, I8, I12, I13, I15, I16	DNN	0.65	Analyzed spectral measurements to improving SOC prediction accuracy by including soil moisture and laboratory indices.	1. Only considers multispectral data. 2. New indices, not developed.
Dhiman et al. [198]	MS	I1, I2, I5, I8, I13, I17	GB	0.97	Estimated soil nutrients and textures, revealing that raw data proved more effective than atmospheric correction.	1. Only considers multispectral data. 2. New indices, not developed.
Babbar et al. [50]	MS	I1, I2, I8, I9, I10, I11, I12, I13	Multilinear Regression analysis	0.97	Finding significant correlations between soil nutrients and various spatial indices.	1. Only considers multispectral data. 2. New indices, not developed.
John et al. [326]	MS	I1, I4, I12, I13, I17, I18, I19, I20	RF	0.68	Indicating the importance of soil nutrient indicators and topographic features in spatial prediction.	1. Only considers multispectral data. 2. New indices, not developed.
Kumar et al. [327]	MS	I1, I21, I22, I23, I24	Regression Kriging	0.53	Regression kriging mapped SOC using satellite-derived variables, presenting correlations with indices like NDVI.	1. Only considers multispectral data. 2. New indices, not developed.
Kalambukattu et al. [63]	MS	12, I8, I21, I25	ANN	0.83	Employed ANN and satellite data to map soil nutrients.	1. Only considers multispectral data. 2. New indices, not developed.

Note: I1-NDVI, I2-EVI, I3-SAVI, I4-RVI, I5-NDWI, I6-RI, I7-CI, I8-ARVI, I9-NPCRI, I10-RDVI, I11-SCI, I12-NDMI, I13-SAVI, I14-NDSI, I15-BI, I16-BSI, I17-NDBI, I18-Clay Index, I19-Iron Index, I20-LST, I21-VTCI, I22-BI, I23-WI, I24-GI, I25-HI, (HS-Hyperspectral, MS-Multispectral)

Table 7.2: Previous studies on soil pH and SOC prediction using AfSIS [3] and LUCAS-2009 [1]

Previous Work	**Dataset**	**Soil Property**	**Model**	**Results**
Zhao et al. [189]	AfSIS, LUCAS-2009	pH, SOC	CNN	pH: R^2=0.81, RMSE=0.41 (AfSIS); R^2=0.85, RMSE=0.53 (LUCAS-2009) SOC: R^2=0.936, RMSE=0.29 (AfSIS); R^2=0.934, RMSE=23.72 (LUCAS-2009)
Akande et al. [354]	AfSIS	pH, SOC	SVR	pH: R^2=0.84, RMSE=0.36; SOC: R^2=0.88, RMSE=0.391
Zhao et al. [190]	AfSIS, LUCAS-2009	SOC	Squeeze and Excitation CNN	R^2=0.83, RMSE=0.276 (AfSIS); R^2=0.935, RMSE=23.565 (LUCAS-2009)
Zhao et al. [65]	AfSIS, LUCAS-2009	SOC	CNN with Quantile Regression	R^2=0.92, RMSE=0.301 (AfSIS); R^2=0.93, RMSE=26.667 (LUCAS-2009)
Zhong et al. [191]	LUCAS-2009	pH, SOC	DCNN	pH: R^2=0.93, RMSE=0.33; SOC: R^2=0.95, RMSE=19.837

7.2, primarily employed ML and DL models using raw spectral bands as input [65, 189–191, 354]. However, none have investigated the role of spectral indices in these datasets for nutrient prediction. Hence, this thesis emphasizes the development of novel spectral indices derived from hyperspectral data, aiming to enhance prediction accuracy while reducing computational complexity compared to conventional ML/DL approaches.

7.3 Experiment Methodology

The methodology employed in this work is outlined below, comprising a series of essential steps designed to ensure a comprehensive investigation and comparison of spectral indices. Figure 7.1 shows the two discrete objectives, each symbolizing a critical component of the approach. These objectives collectively guide the optimization process, ensuring a balanced performance across various aspects of the model. The detailed steps of these two objectives are as follows:

1. The first objective begins with the process that involves reading, cleaning, and normalizing the data.

2. Then, spectral bands and various preexisting spectral indices related to vegetation,

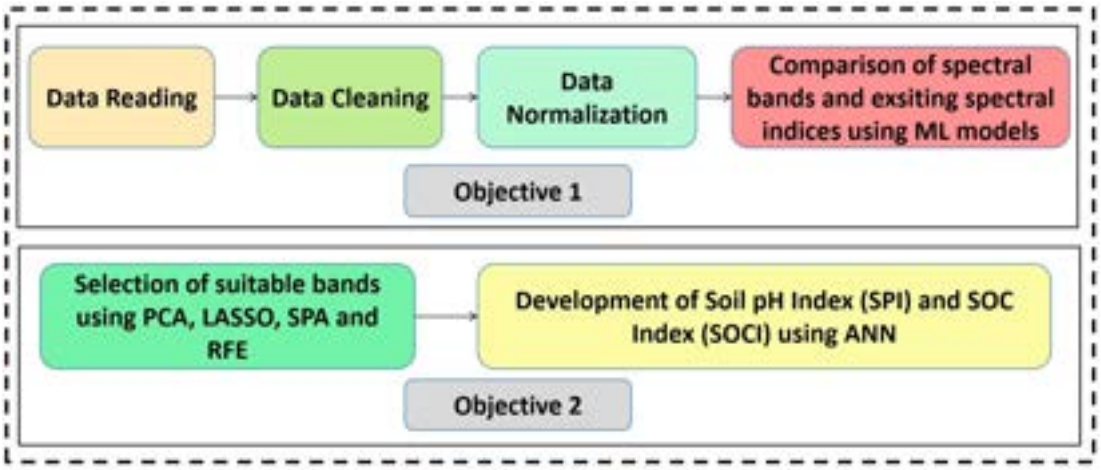

Figure 7.1: Research methodology flowchart for the development of new spectral indices

soil, water, temperature, and thermal properties are analyzed using several ML models to assess performance in different environmental conditions.

3. Employ feature selection techniques for PCA, LASSO, SPA, and RFE to identify key spectral bands and compare the advantages for reliable prediction.

4. Subsequently, use an ANN to create novel spectral indicators designed explicitly for soil pH and SOC.

7.3.1 Dataset Description

This thesis confirms the effectiveness of the proposed approach by employing AfSIS [3] and LUCAS-2009 [1] as validation sources. These datasets are foundational hyperspectral resources supporting large-scale soil mapping efforts across Africa and Europe. Such datasets provide point-based data, which is essential for assessing soil properties. With a comprehensive range of spectral signatures across various wavelengths, these datasets enable detailed analysis of soil composition by capturing subtle differences in reflectance patterns, which reflect variations in nutrient levels and soil structure. AfSIS [3] and LUCAS-2009 [1] are valued for the high-quality imaging spectroscopic data and broad geographic coverage, offering extensive support for sustainable land management research—the detailed descriptions of these datasets are mentioned in Chapter 3.

7.3.2 Data Preprocessing

This thesis applies comprehensive preprocessing to the hyperspectral data in the Comma Separated Value (CSV) format, a critical step to ensure the accuracy and reliability of subsequent analyses. Data preprocessing is vital as it transforms raw data into a clean and consistent format, making it suitable for further analysis and ML applications [154]. For this work, data preprocessing includes cleaning and normalizing the data.

7.3.2.1 Data Cleaning

Initially, the process addresses any inconsistencies, missing values, and outliers through data cleaning. This step involved several sub-processes: removing rows or columns with excessive missing values, imputing missing data points using appropriate statistical methods, correcting erroneous entries, and standardizing the format to maintain uniformity. Data cleaning ensures the dataset is free from anomalies that could distort the analytical results [356].

7.3.2.2 Data Normalization

Following data cleaning, the dataset is normalized using the *Standard Scaler* [1]. This technique normalizes the averages and standard deviations of the characteristics to 0 and 1, respectively [357]. Ensuring consistent scales is crucial for the algorithm to efficiently learn patterns and correlations in hyperspectral data, where each feature corresponds to a distinct wavelength band. The normalization method enhances the overall stability and effectiveness of the model in managing hyperspectral datasets. Equation 7.1 demonstrates the calculation of the standard scalar.

$$z = \frac{x - \mu}{\sigma} \tag{7.1}$$

In this context, z represents the scaled data, x represents the data that needs to be scaled, μ represents the mean of the training samples, and σ represents the standard deviation of the training samples.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

7.3.3 Experimental Framework

This work presents the experimental framework, which consists of two objectives, as depicted in Figure 7.1. These objectives guide the design and evaluation of the proposed methodology throughout the various stages of the study.

7.3.3.1 Objective 1: Comparison of Spectral Bands and Existing Spectral Indices with ML Models

This work's first objective is to assess each spectral indicator's efficacy in various ML algorithms, emphasizing the advantages and constraints of each method in forecasting soil pH and SOC. Before proposing new spectral indices, it is crucial to compare spectral bands with existing spectral indices for predicting soil pH and SOC. It establishes a baseline and identifies current limitations. Figure 7.2 illustrates the flowchart for comparing spectral bands and existing spectral indices, which begins with data preprocessing, followed by the computation and categorization of existing indices (VI, SI, WI, TTI, and all indices combined). This work compares the performance of ML models for each category to predict soil pH and SOC. This thorough evaluation benchmarks current indices, highlights the strengths and weaknesses, and guides the development of more effective new indices.

The list of ML models used to assess these spectral indices for soil pH and SOC prediction is as follows:

- ***Partial Least Square regression (PLSR):*** It is a multivariate regression model for identifying the underlying relationships between two matrices, specifically the predictor and response variables [92].
- ***Support Vector Regression (SVR):*** It is an ML model that is capable of performing both linear and non-linear regression [354].
- ***Artificial Neural Network (ANN):*** It is a DL model that works like the human brain. It can capture intricate correlations between inputs and outputs [63].
- ***Genetic Expression Programming (GEP):*** It is an evolutionary approach that uses computer programs or expressions to model data effectively [358].

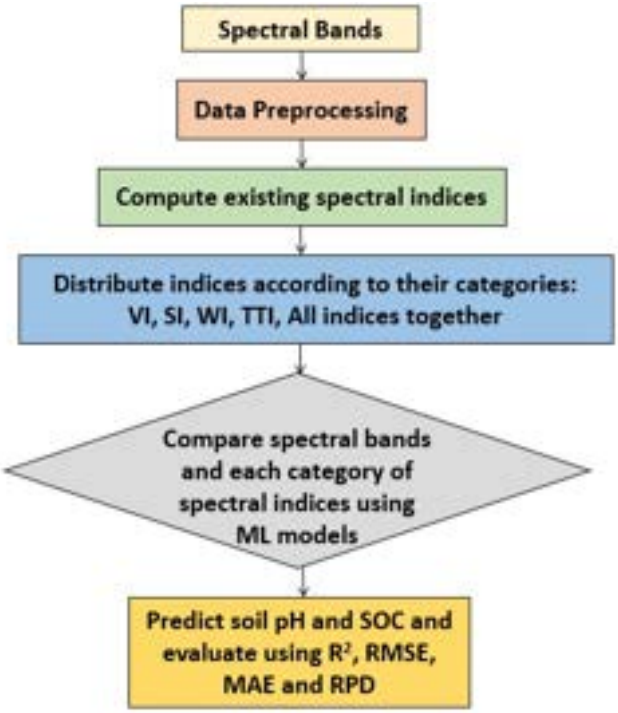

Figure 7.2: Flowchart for comparing spectral bands and existing spectral indices (VI- vegetation indices, SI- soil indices, WI- water indices, TTI- temperature and thermal indices)

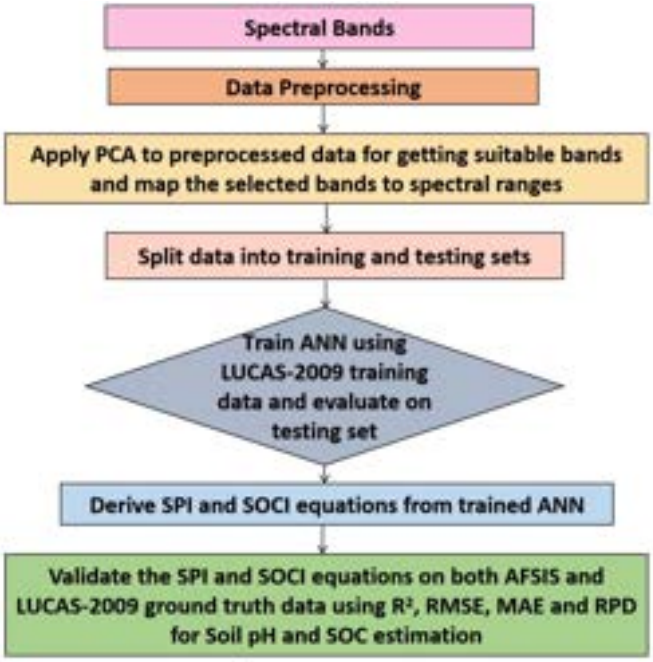

Figure 7.3: Flowchart for Design and Validation of New Soil Spectral Indices

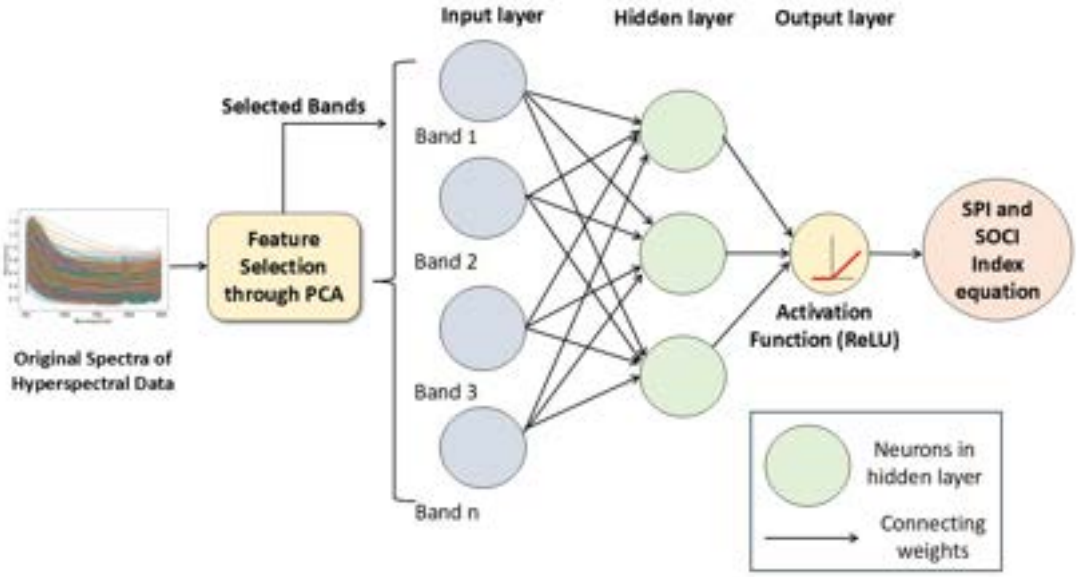

Figure 7.4: Architecture to Compute SPI and SOCI Indices using PCA and ANN

7.3.3.2 Objective 2: Development of a New Spectral Index for Soil pH and SOC

Developing a new spectral index for soil pH and SOC involves a two-step approach utilizing PCA and ANN. Initially, it employs multiple feature selection techniques, including PCA, LASSO, SPA, and RFE. The results in Tables 7.6 and 7.7 indicate that PCA gives the optimal spectral bands from a large dataset, identifying those most relevant for predicting soil properties. This dimensionality reduction technique ensures the use of only the most informative spectral bands, enhancing the efficiency and accuracy of the analysis. Subsequently, this work applies an ANN on these selected spectral bands to create a new spectral index. Training the ANN on the optimal features enables the development of a model capable of capturing complex, non-linear relationships between the spectral data and soil pH and SOC. This combined approach of PCA and ANN yields a robust and precise spectral index, significantly enhancing the accuracy of predicting soil properties and facilitating the effective soil management practices.

Figure 7.3 illustrates the flowchart for designing and validating the proposed SPI and SOCI indices to estimate soil pH and SOC using an ANN. The method begins by preprocessing the reflectance data and then utilizing PCA to identify and pick the most informative bands. Subsequently, it divides the data into training and testing sets. The training dataset trains the ANN. It evaluates the model using a separate testing set and

Algorithm 1: SPI and SOCI Indices Construction

Input: Hyperspectral reflectance data D, soil pH, and SOC.
Output: SPI and SOCI equations.
1. Preprocess reflectance data D
2. Apply PCA to reflectance data (select informative Bands)
   $B_1, B_2, ..., B_m \leftarrow PCA(D)$
3. Map selected Bands to spectral ranges:
   Bands ← Mapping into spectral ranges (e.g.,Red, NIR, SWIR, MIR)
4. Split data:
   $D_{train}, D_{test} \leftarrow$ Split of D in ratio 7:3
5. Train ANN on LUCAS-2009 data:
   $W,b \leftarrow$ Train ANN (D_{train}, pH/SOC)
6. Derive SPI and SOCI equations:
   $SPI = f_{ANN}(B_1, B_2, ..., B_m)$
   $SOCI = f_{ANN}(B_1, B_2, ..., B_m)$
7. Validate SPI and SOCI indices using AfSIS and LUCAS-2009 ground truth data and evaluate using R^2, RMSE, MAE and RPD

then uses the results to construct equations for the SPI and SOCI indices. The proposed indices equations are validated on AfSIS and LUCAS-2009 ground truth data using evaluation metrics.

Figure 7.4 depicts the architecture for computing proposed SPI and SOCI indices using PCA and ANN. Initially, the hyperspectral data capture detailed reflectance values across a range of wavelengths. PCA reduces dimensionality by selecting the most informative bands. These bands are then input into an ANN with ReLU activation to learn complex data relationships, ultimately calculating the SPI and SOCI indices for improved predictions of soil properties.

Algorithm 1 presents a streamlined workflow for developing the SPI and SOCI indices from hyperspectral reflectance data. The process begins with preprocessing the spectra and applying PCA to extract the most informative bands, which are then mapped to relevant spectral regions for interpretability. The dataset is split into training and testing subsets, and an ANN is trained on the LUCAS-2009 data [1] to learn the relationship between the selected bands and soil properties. The trained model is used to derive the SPI and SOCI equations, which are subsequently validated using AfSIS [3] and LUCAS-2009 [1] ground-truth data based on R^2, RMSE, MAE, and RPD, ensuring reliable and reproducible index development.

7.3.3.3 Multiple Feature Selection Techniques

This work describes various feature selection techniques for identifying suitable bands for the SPI and SOCI.

1. ***Principal Component Analysis (PCA):*** The principal components effectively capture the highest variation in the data, reducing the number of dimensions and identifying the most important spectral bands [359].

2. ***Least Absolute Shrinkage and Selection Operator (LASSO):*** It is a regression technique that adds a penalty term to shrink some feature coefficients to zero, efficiently selecting relevant features and reducing model complexity [360].

3. ***Successive Projection Algorithm (SPA):*** It guarantees that the desired characteristics have minimal correlation, resulting in a varied and valuable collection of bands [361].

4. ***Recursive Feature Elimination (RFE):*** RFE determines the most essential features for a prediction job by repeatedly training the model and removing the least impactful features [362].

7.4 Results

7.4.1 Objective 1: Assessment of Existing Spectral Indices

Spectral indices are the mathematical combinations of various spectral bands, which help to enhance specific features and provide more meaningful inputs compared to raw spectral bands. These indices include VI, SI, WI, TTI, and a comprehensive set of all indices. Table 7.3 and 7.4 focus on the predictive performance of various spectral bands and existing spectral indices for soil pH and SOC prediction across different datasets (AfSIS [3] and LUCAS-2009 [1]) using four ML models: PLSR, SVR, ANN, and GEP. It provides a comparative overview, highlighting which spectral features and models yield the most accurate predictions for each soil property. Table 7.5 shows the selected hyperparameters for each ML model. These hyperparameters are optimized

Table 7.3: Results of spectral bands and existing spectral indices used as input in ML models for soil pH prediction

Soil pH (AfSIS)					
Input Category	**Model**	**R^2**	**RMSE**	**MAE**	**RPD**
Spectral bands	**ANN**	**0.8**	**0.42**	**0.3**	**2.26**
Vegetation Indices	ANN	0.52	0.66	0.51	1.45
Soil Indices	ANN	0.57	0.62	0.47	1.53
Water Indices	PLSR	-0.001	0.96	0.73	0.99
Temperature and Thermal Indices	ANN	0.17	0.87	0.66	1.1
All indices	ANN	0.59	0.6	0.47	1.58
[92] (A subset of vegetation indices VI1, VI2, VI3, VI4, VI5)	PLSR	0.04	0.93	0.73	1.02
Soil pH (LUCAS-2009)					
Input Category	**Model**	**R^2**	**RMSE**	**MAE**	**RPD**
Spectral bands	**ANN**	**0.87**	**0.46**	**0.35**	**2.88**
Vegetation Indices	SVR	0.45	1	0.77	1.35
Soil Indices	ANN	0.38	1.05	0.87	1.27
Water Indices	PLSR	0.095	1.28	1.1	1.05
Temperature and Thermal Indices	ANN	0.26	1.15	0.96	1.16
All indices	ANN	0.48	0.96	0.47	1.39
[92] (A subset of vegetation indices VI1, VI2, VI3, VI4, VI5)	PLSR	0.07	1.3	1.11	1.03

Note: VI1-NDVI, VI2-EVI, VI3-RVI, VI4-NDWI, VI5-ARVI

to achieve the best predictive performance for soil nutrient prediction. It also serves as a guide for future modeling efforts using similar datasets.

The ANN model using spectral bands shows superior performance in all input categories for soil pH and SOC prediction as shown in Table 7.3 and 7.4. The input category of all indices best predicted AfSIS soil pH with ANN, achieving R^2 of 0.59, indicating its strong predictive power when utilizing comprehensive spectral information. Similarly, for the LUCAS-2009 soil pH, the ANN model with all indices achieved the highest performance (R^2 = 0.48), underscoring the importance of broad spectral data in capturing variations in soil pH. For AfSIS SOC, the ANN model with soil indices again shows the highest success, with R^2 = 0.77, suggesting the effectiveness of soil-specific indices in specific contexts. The GEP model excelled in predicting SOC for the LUCAS-2009 dataset with R^2 = 0.72, again demonstrating the value of

Table 7.4: Results of spectral bands and existing spectral indices used as input in ML models for SOC prediction

SOC (AfSIS)					
Input Category	**Model**	**R^2**	**RMSE**	**MAE**	**RPD**
Spectral bands	**ANN**	**0.94**	**0.33**	**0.26**	**3.8**
Vegetation Indices	SVR	0.74	0.61	0.38	1.98
Soil Indices	ANN	0.77	0.58	0.35	2.11
Water Indices	ANN	0.01	1.21	0.83	1
Temperature and Thermal Indices	ANN	0.2	1.09	0.73	1.11
All indices	ANN	0.7	0.67	0.4	1.82
[92] (A subset of vegetation indices VI1, VI2, VI3, VI4, VI5)	PLSR	0.31	1.01	0.7	1.21
SOC (LUCAS-2009)					
Input Category	**Model**	**R^2**	**RMSE**	**MAE**	**RPD**
Spectral bands	**ANN**	**0.90**	**27.82**	**21.8**	**3.43**
Vegetation Indices	ANN	0.24	78.31	41.06	1.15
Soil Indices	ANN	0.24	78.24	38.96	1.15
Water Indices	ANN	0.02	88.8	46.78	1.01
Temperature and Thermal Indices	GEP	0.72	48.01	22.81	1.89
All indices	PLSR	0.3	75.24	37.67	1.19
[92] (A subset of vegetation indices VI1, VI2, VI3, VI4, VI5)	PLSR	0.06	87.05	46.59	1.03

VI1-NDVI, VI2-EVI, VI3-RVI, VI4-NDWI, VI5-ARVI

comprehensive indices. Overall, the input category of all spectral indices, followed by spectral bands, consistently leads to better model performance, highlighting the importance of integrating a wide range of spectral information to enhance the accuracy and reliability of soil property predictions. Results indicate that the ANN model is superior in most cases, particularly when utilizing spectral bands and existing indices for predicting soil pH and SOC.

7.4.2 Objective 2: Screening Results of Multiple Feature Selection Techniques

For more precise and reliable predictions of soil pH and SOC, novel indices that better capture their spectral characteristics can be developed to complement existing spectral bands and indices. While existing approaches have demonstrated reasonable

Table 7.5: Selected hyperparameters of various ML models for spectral indices

Model	**Hyperparamters**
PLSR	n_components = 4, scale = True, max_iter = 100
SVR	kernel = 'rbf', C = 50, gamma = 0.1, epsilon = 0.1
ANN	activation = 'relu', optimizer = 'Adam' , learning_rate = 0.028, epochs = 100, batch_size = 32, verbose = 1
GEP	generations = 50, population_size = 1000, tournament_size=5, const_range = (-2.0, 2.0), function_set = ('add', 'sub', 'mul', 'div', 'sqrt', 'log', 'abs', 'sin', 'cos'), p_crossover = 0.7, p_subtree_mutation = 0.1, p_hoist_mutation = 0.1, p_point_mutation = 0.1, max_samples=0.9, verbose=1

performance in predicting soil nutrients like phosphorus, further improvements may enhance prediction accuracy and reliability across varying soil conditions. For this, feature selection techniques help extract the suitable spectral range for developing novel spectral indices. From objective 1, it is clear that ANN shows the best R^2 in the range of 0.8 to 0.94 for SOC and soil pH. Hence, this work combines multiple feature selection techniques with ANN to get suitable bands for SPI and SOCI. The analysis across Tables 7.6 and 7.7, which showcase various feature selection techniques for predicting soil pH and SOC, reveals notable differences in performance across datasets. For soil pH and SOC prediction using the AfSIS and LUCAS-2009 datasets, PCA with ANN consistently demonstrates a strong correlation (R^2 values ranging from 0.86 to 0.952). However, while PCA excels in capturing variance, LASSO with ANN and RFE with ANN offer comparable results.

In contrast, SPA with ANN consistently emerges as the least effective technique across all datasets. For instance, in soil pH prediction using both datasets, SPA with ANN shows near-zero or negative R^2 values, indicating poor correlation and predictive capability. Moreover, SPA with ANN yields significantly higher RMSE and MAE values than other techniques, indicating its inability to effectively select relevant features. It underscores SPA's unsuitability for these datasets, likely due to its inability to handle the complexities and interactions in soil pH and SOC prediction models. While PCA shows promise for capturing variance, the choice of feature selection technique critically impacts predictive accuracy and model performance, necessitating careful consideration based on specific dataset characteristics and predictive goals.

Table 7.6: Results of various feature selection techniques with ANN for soil pH prediction

Soil pH (AfSIS)				
Feature Selection Technique	**R^2**	**RMSE**	**MAE**	**RPD**
PCA	**0.86**	**0.4**	**0.26**	**2.39**
LASSO	0.67	0.55	0.4	1.74
SPA	-0.003	0.96	0.72	0.99
RFE	0.6	0.6	0.43	1.59
Soil pH (LUCAS-2009)				
Feature Selection Technique	**R^2**	**RMSE**	**MAE**	**RPD**
PCA	**0.945**	**0.42**	**0.31**	**3.19**
LASSO	0.34	1.09	0.91	1.23
SPA	0.11	1.26	1.1	1.06
RFE	0.54	0.98	0.76	1.87

Table 7.7: Results of various feature selection techniques with ANN for SOC prediction

SOC (AfSIS)				
Feature Selection Technique	**R^2**	**RMSE**	**MAE**	**RPD**
PCA	**0.952**	**0.28**	**0.18**	**4.69**
LASSO	0.93	0.31	0.22	3.84
SPA	0.144	1.13	0.73	1.08
RFE	0.932	0.31	0.209	3.85
SOC (LUCAS-2009)				
Feature Selection Technique	**R^2**	**RMSE**	**MAE**	**RPD**
PCA	**0.943**	**19.62**	**10.41**	**4.59**
LASSO	0.87	31.93	18.89	2.82
SPA	0.62	54.92	28.71	1.64
RFE	0.45	67.99	35.1	1.08

7.4.3 Objective 2: Index-Based Analysis of Soil pH and SOC

An index is a numerical value derived from a mathematical formula that combines information from multiple spectral bands of a spectral image. Sensors detect electromagnetic events in terms of bands, which correspond to specific wavelengths of light. Band calculations yield unique results that facilitate the detection of the desired signature of an element or material. Such calculations are used here for band analysis and further prediction of soil pH and SOC indices.

PCA selects five bands for soil pH and eight bands for SOC due to the specific spectral sensitivities of these properties. It reduces data dimensionality and highlights critical spectral features related to soil properties. While researchers have explored

various bands for these parameters, this selection yields the best results, reflecting the optimal balance between capturing essential spectral information and achieving accurate predictions of soil pH and SOC.

This thesis constructs various equations to generate the SPI and SOCI indices using the PLSR, SVR, ANN, and GEP models. Through rigorous testing and validation, ANN consistently outperforms other models, delivering the most precise estimates of soil pH and SOC. While techniques like PLSR and SVR offer fast training, this thesis chooses ANN due to its superior ability to capture complex relationships within the data. ANN is employed to derive equations for the proposed spectral indices, leveraging its ability to model complex non-linear relationships. This focus on accuracy proved fruitful, as ANNs yielded significantly higher prediction results for both soil pH and SOC, aligning perfectly with the research goals of achieving the most precise estimates of soil properties.

The reported metrics, as mentioned in the chapter 3, prioritize the index's ability to predict soil properties on entirely new data. This focus on generalizability is essential for real-world applications. A common 70:30 split is used to separate the data for calibration and validation. This approach ensures robust model training while reserving a significant portion for unbiased evaluation and validation. Consequently, the testing R^2 becomes the key metric for assessing generalizability.

The equations of the new index for soil pH and SOC are defined in equations 7.2 and 7.3 respectively:

$$\begin{aligned}\text{Soil pH Index (SPI)} = W_1 \cdot B(\text{MidIR1}) + W_2 \cdot B(\text{SWIR1}) + W_3 \cdot B(\text{Red}) - \\ W_4 \cdot B(\text{MidIR2}) - W_5 \cdot B(\text{SWIR2}) - b\end{aligned} \tag{7.2}$$

Here W_1 = 0.19, W_2 = 0.48, W_3 = 0.15, W_4 = 0.08, W_5 = 0.06 and b = 0.04. The B(MidIR1) represent the MidInfrared Band 1, B(MidIR2) represent the MidInfrared Band 2, B(SWIR1) represents the Short-wave Infrared Band 1, B(SWIR2) represents

Table 7.8: Comparison of proposed SPI index with previous work on soil pH prediction using spectral bands

Soil pH (AfSIS)					
Reference	**Technique**	**R^2**	**RMSE**	**MAE**	**RPD**
[189]	CNN with spectral bands	0.81	0.41	-	-
[354]	SVR with spectral bands	0.84	0.36	0.25	-
This thesis	**Proposed SPI Index**	**0.86**	**0.4**	**0.26**	**2.39**
Soil pH (LUCAS-2009)					
Reference	**Technique**	**R^2**	**RMSE**	**MAE**	**RPD**
[189]	CNN with spectral bands	0.85	0.536	-	-
[191]	DCNN with spectral bands	0.93	0.334	-	-
This thesis	**Proposed SPI Index**	**0.945**	**0.42**	**0.31**	**3.19**

Short-wave Infrared Band 2 and B(Red) represents the Red Band.

$$\begin{aligned} \text{SOC Index (SOCI)} = W_1 \cdot B(\text{MidIR1}) - W_2 \cdot B(\text{SWIR1}) - W_3 \cdot B(\text{NIR}) - \\ W_4 \cdot B(\text{MidIR2}) - W_5 \cdot B(\text{SWIR2}) + W_6 \cdot B(\text{SWIR3}) - W_7 \cdot B(\text{SWIR4}) + \\ W_8 \cdot B(\text{SWIR5}) - b \end{aligned} \tag{7.3}$$

Here $W_1 = 0.20$, $W_2 = 0.27$, $W_3 = 0.08$, $W_4 = 0.21$, $W_5 = 0.15$, $W_6 = 0.28$, $W_7 = 0.008$, $W_8 = 0.13$, and b = 0.001. The B(MidIR1) represent the MidInfrared Band 1, B(MidIR2) represent the MidInfrared Band 2, B(SWIR1) represents the Short-wave Infrared Band 1, B(SWIR2) represents Short-wave Infrared Band 2, B(SWIR3) represents Short-wave Infrared Band 3, B(SWIR4) represents Short-wave Infrared Band 4, B(SWIR5) represents Short-wave Infrared Band 5, and B(NIR) represents the Near-Infrared Band. The parameter B represents the reflectance values at specific wavelengths, and the coefficients of bands are the weights that the ANN model determines. The constant value at the end is the random error term. This approach enhances the predictive accuracy of specific soil nutrients, enabling more precise and reliable assessments of soil health.

The SPI and SOCI indices differ from hyperspectral band predictions in the approach to utilizing spectral data. This thesis specifically designed the SPI index to predict soil pH by integrating spectral information into a single index that correlates strongly with soil acidity or alkalinity. Similarly, the SOCI index estimates SOC by integrating specific spectral features associated with organic carbon. In contrast, hyperspectral

Table 7.9: Comparison of proposed SOCI index with previous work on SOC prediction using spectral bands

SOC (AfSIS)					
Reference	**Technique**	**R^2**	**RMSE**	**MAE**	**RPD**
[189]	CNN with spectral bands	0.936	0.29	-	-
[190]	Squeeze and excitation CNN with spectral bands	0.83	0.276	-	-
[354]	SVR with spectral bands	0.88	0.391	0.162	-
[65]	CNN with quantile regression with spectral bands	0.92	0.301	-	-
This thesis	**Proposed SOCI Index**	**0.952**	**0.28**	**0.18**	**4.69**
SOC (LUCAS-2009)					
Reference	**Technique**	**R^2**	**RMSE**	**MAE**	**RPD**
[189]	CNN with spectral bands	0.93	23.722	-	-
[190]	Squeeze and excitation CNN with spectral bands	0.935	23.565	-	-
[65]	CNN with quantile regression with spectral bands	0.93	26.667	-	-
[363]	DCNN with spectral bands	0.95	19.837	-	-
This thesis	**Proposed SOCI Index**	**0.963**	**19.62**	**10.41**	**4.59**

band predictions involve using ML models to directly analyze the full spectrum of bands, capturing complex relationships between spectral data and soil properties. SPI and SOCI offer several advantages over direct hyperspectral predictions. First, these indices simplify the interpretation of spectral data by reducing it to a single, focused index for predicting soil pH and SOC, respectively. It reduces computational complexity and resource requirements compared to analyzing the full hyperspectral spectrum. Additionally, SPI and SOCI are specialized indices that can enhance prediction accuracy for the specific targets, making them easier to use.

The analysis of Table 7.8 and 7.9 highlights the impressive performance of the proposed SPI and SOCI indices to predict soil pH and SOC. For AfSIS soil pH, the proposed SPI index achieves R^2 of 0.86. For LUCAS-2009 soil pH, the results are even more impressive with an R^2 of 0.945. For SOC prediction, the proposed SOCI index achieves R^2 of 0.952 for the AfSIS dataset and R^2 of 0.963 for the LUCAS-2009 dataset. The findings demonstrate that the proposed SPI and SOCI indices, tailored to specific soil characteristics (soil pH and SOC), yield superior outcomes. Thus, this work determines that developing new indices for specific soil properties is crucial.

Figure 7.5 illustrates the proposed SPI and SOCI indices training and testing RMSE

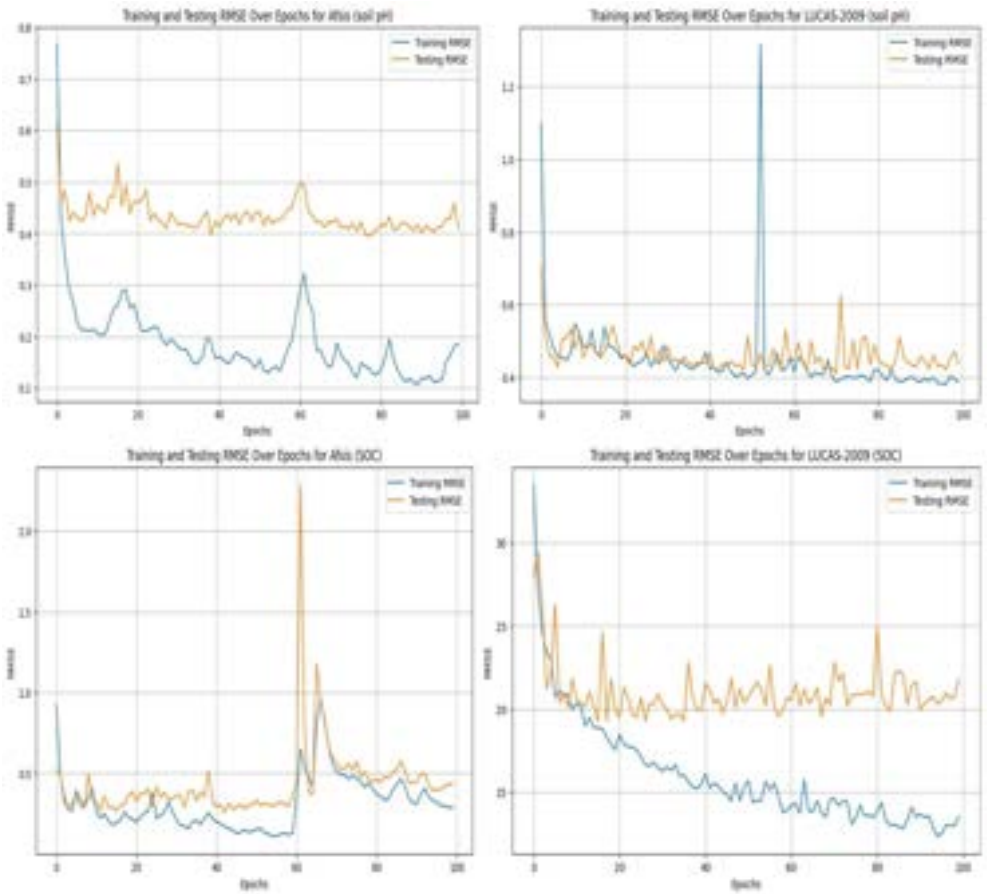

Figure 7.5: Proposed SPI and SOCI indices training & testing error for soil pH and SOC prediction

over epochs for soil pH and SOC prediction using the AfSIS and LUCAS-2009 datasets. Both datasets show a significant reduction in training RMSE over the initial epochs for soil pH prediction using the SPI index, indicating its effective learning. The testing RMSE maintains a steady trend with slight fluctuations, demonstrating the SPI index's ability to generalize well to new data. This consistent performance reflects the robustness of the SPI index in predicting soil pH across various time periods. For SOC prediction using the SOCI index, the training RMSE decreases significantly in the early epochs, showcasing its learning capability. The testing RMSE follows a stable trend, with minor variations, indicating reliable performance in predicting SOC. The models demonstrate the effectiveness and consistency in generalizing SOC predictions across various epochs, highlighting the robustness and reliability.

Figure 7.6 shows the proposed SPI and SOCI indices residual plot for soil pH and SOC prediction using AfSIS and LUCAS-2009 datasets. The residuals are scattered

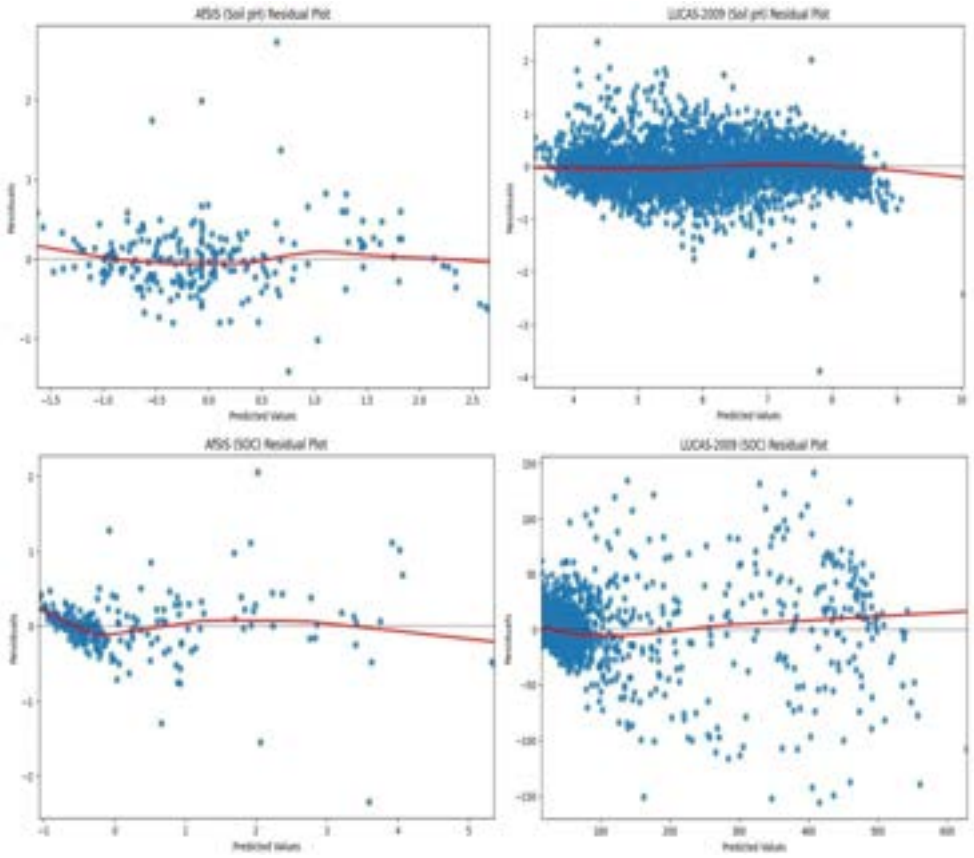

Figure 7.6: Proposed SPI and SOCI indices residual plot for soil pH and SOC prediction

around the horizontal axis, indicating relatively unbiased predictions. However, a slight pattern where residuals increase and decrease suggests minor non-linearity. For AfSIS soil pH, most residuals are within the range of -1 to 2, demonstrating that the SPI index generally predicts soil pH accurately. In the LUCAS-2009 soil pH residual plot, residuals are more tightly clustered around the horizontal axis, indicating fewer extreme errors and more consistent SPI index performance. For SOC prediction using the SOCI index, the residual plots reveal residuals fairly dispersed around the horizontal axis, with some evident non-linear patterns.

7.5 Discussion

7.5.1 Advantages of Proposed Spectral Indices

1. ***Cost and Time Effective:*** The use of SPI and SOCI reduces the need for extensive soil sampling and laboratory testing, making the process more efficient and less

expensive.

2. ***Improved Accuracy:*** The SPI and SOCI indices are designed for specific wavelengths that best reflect variations in soil pH and SOC. Unlike existing indices, which may broadly capture soil characteristics, SPI and SOCI specifically target the spectral bands most sensitive to pH and SOC levels. This targeted approach enhances prediction accuracy by focusing on wavelengths with the strongest correlations to specific soil properties, thereby improving overall performance in predicting soil properties.

3. ***Enhanced Agricultural Decision-Making:*** Using SPI and SOCI indices enables precise assessment of soil pH and SOC, guiding better agricultural decisions. SPI indicates the optimal pH for nutrient availability, while SOCI helps determine SOC levels, which are crucial for soil health. Together, these indices help farmers fine-tune fertilizer to improve crop yield, enhance soil fertility, and support sustainable farming practices.

4. ***Reduced Computational Complexity:*** By simplifying the feature space and focusing on relevant wavelengths, the indices reduce computational load, making them efficient for large-scale soil monitoring tasks. Ultimately, this contributes to better soil health and crop productivity, demonstrating the significant impact of using tailored spectral indices in soil nutrient analysis.

7.5.2 Limitations of Proposed Spectral Indices

1. ***Limited to Soil pH and SOC:*** SPI and SOCI indices are optimized to predict only soil pH and SOC, limiting the utility in predicting other soil nutrients or properties. The wavelength selection and formulation are tailored to the spectral characteristics of these specific targets, which may not effectively capture the spectral responses associated with other soil constituents.

2. ***Generalizability Concerns:*** While effective in this chapter, the generalizability of these indices to diverse geographical, environmental, and sensor conditions remains a potential limitation. Variations in soil type, texture, moisture, mineralogy, and differences in spectrometer characteristics such as wavelength

range, resolution, and calibration can significantly alter spectral responses, potentially reducing the robustness and transferability of these indices across datasets and instruments.

7.6 Summary

This chapter compares multiple feature selection techniques and explores, analyzes, and develops spectral indices for predicting soil nutrients. It utilises hyperspectral data to evaluate a comprehensive range of existing indices and reveals the limited effectiveness in accurately predicting soil pH and SOC. This thesis develops two novel indices, SPI and SOCI, utilizing PCA and ANN techniques to significantly enhance prediction performance. It emphasises the importance of developing soil-specific spectral indices and demonstrates that while multispectral data is operationally practical, hyperspectral data offers the precision needed for robust modelling.

This chapter is based on the following publication:

Journal 2: S. Jain, D. Sethia and K.C. Tiwari. "Developing Novel Spectral Indices for Precise Estimation of Soil pH and Organic Carbon With Hyperspectral Data and ML." **Environmental Monitoring and Assessment**: 196, 1255 (November 2024). (**SCIE, Impact factor-3.6, Publisher: Springer**).

Chapter 8

Conclusion, Future Work & Social Impact

This chapter summarizes the thesis's key findings and contributions, reflecting on how the overall objectives have been addressed. It discusses the implications of these results both academically and in practical, real-world applications. The chapter also highlights potential directions for future research, identifying areas where further work could refine existing methods or explore new avenues of investigation. Additionally, it considers the societal impact of the study, illustrating how advances in soil nutrient prediction through spectral data can benefit agriculture, environmental monitoring, and sustainable land management.

8.1 Conclusion

This thesis aims to improve the accuracy and scalability of soil nutrient prediction using advanced ML/DL and spectral analysis techniques. It presents the four research objectives, each addressed through multiple experiments, datasets, and peer-reviewed publications.

The first objective (RO1) is to critically review the current state of soil nutrient prediction using spectral data and artificial intelligence. This objective reveals that while many studies have applied ML or DL models to hyperspectral or multispectral data, these models often lack robust feature selection strategies or spectral indices

specifically tailored to soil nutrient prediction. Most existing indices are designed for assessing vegetation or water content, rather than soil nutrients. It is covered in chapter 3.

The second objective (RO2) is addressed through the development of the novel spectral indices. The novel indices integrate multiple feature selection techniques before feeding the selected features into an ANN model. This method significantly reduced the dimensionality of hyperspectral data, removed redundant wavelengths, improved prediction accuracy, and reduced overfitting. It is covered in chapter 7.

The third objective (RO3) involves comparing the performance of various ML and DL models for soil nutrient prediction. It is achieved by proposing a hybrid ML model (HyPLS-GPR) and two DL-based frameworks (CNViS-Net and CATMTL-Net). A HyPLS-GPR using a hybrid PLSR-GPR model demonstrates reliable performance in predicting multiple soil nutrients from multispectral data. On the other hand, the CNViS-Net framework, which combines CNN and ViT for feature extraction, followed by SVR for final regression, shows the best performance on hyperspectral data. Similarly, the CATMTL-Net framework, which integrates convolutional and attention-based multitarget learning, achieves superior performance in predicting multiple soil nutrients from hyperspectral data. These comparisons demonstrate that hyperspectral data, paired with DL, achieves higher precision, while multispectral data, combined with lightweight ML models, ensures practicality and scalability. It is covered in chapters 4, 5, and 6.

The fourth and final objective (RO4) focuses on exploring and evolving spectral indices for soil nutrient prediction. Existing spectral indices (vegetation, soil, water, temperature, and thermal) are analyzed for the usefulness in predicting soil pH and SOC. The performance of these indices is limited, especially when applied to multispectral data. This thesis proposes two novel spectral indices to address this issue: the Soil pH Index (SPI) and the Soil Organic Carbon Index (SOCI). These indices are developed using PCA-based band selection and ANN modelling, and achieve higher prediction accuracy than all existing indices across the AfSIS and LUCAS-2009 datasets. It is covered in chapter 7.

8.1.1 Key Insights from This Research

This thesis yields several key learnings about how spectral data can be transformed into accurate and interpretable soil nutrient prediction systems when coupled with appropriate AI models and pre-processing strategies. One of the significant insights is that raw spectral bands alone often contain redundant and noisy information, which reduces the effectiveness of even complex ML models. Including carefully selected wavelengths and engineered spectral indices significantly improves model performance and generalizability.

It is also observed that the choice of model must align with the nature of the spectral data. DL models perform exceptionally well with high-resolution hyperspectral data, capturing complex non-linear patterns in the reflectance curves. However, traditional ML models like SVR, PLSR, and GPR proved more stable and computationally efficient for low-resolution multispectral data. Another important insight is the domain gap between spectral index development in soil applications. Existing indices designed for vegetation monitoring (such as NDVI or SAVI) or soil, water, temperature, or thermal measurements often underperform in soil-nutrient prediction tasks, reaffirming the need for domain-focused index engineering.

Finally, the cross-dataset evaluation shows that combining data-driven feature selection with physical knowledge (e.g., wavelength absorption behaviour) leads to more robust and transferable solutions. The proposed SPI and SOCI indices serve as a starting point for creating interpretable and scalable models for soil health monitoring using AI and remote sensing.

8.1.2 Limitations

This thesis encounters several limitations despite the successful outcomes, which are:

1. ***Incomplete nutrient coverage:*** Although this thesis aims at soil nutrient prediction, only a selected subset of nutrients—primarily soil pH, SOC, Nitrogen, Phosphorus, Potassium, Calcium, and Magnesium—are considered in the modelling stages. It is due to the limited availability of ground-truth nutrient labels in public datasets (Hyperview [4], Rupnagar dataset [5], and Univalle [6]). Other essential nutrients, such as sulfur, iron, and zinc, are either inconsistently

reported or completely missing, which restricts the scope of nutrient-level prediction.

2. ***Lack of India-specific soil spectral open datasets:*** The absence of large-scale, standardized national datasets for Indian soils limits the regional representativeness of the models. Most public datasets (LUCAS-2009 [1], LUCAS-2015 [2], AfSIS [3], HYPERVIEW dataset [4], Univalle dataset [6]) used in this research are from non-Indian contexts, which may not fully capture the diverse soil characteristics, climatic variability, and management practices across Indian agro-ecological zones. Only the Rupnagar dataset [5] represents Indian soil conditions; however, its limited geographic scope restricts its nationwide applicability. Consequently, the direct transferability of the models to Indian conditions remains uncertain until these models are validated on locally acquired spectral datasets.

3. ***Field validation and operational deployment not conducted:*** While the models are thoroughly validated using multiple public datasets, no field-level validation or real-world deployment is performed. Ground-truthing of predictions in operational agricultural conditions or real-time validation using UAV/drone-based sensors would be required before practical adoption. Additionally, integration with user interfaces or farmer advisory systems is beyond the scope of this work.

8.1.3 Future Work

Based on the insights gained and the limitations observed in this research, several promising future directions emerge that can significantly enhance the robustness, scalability, and practical applicability of spectral-based soil nutrient prediction frameworks.

1. ***Extension of spectral indices to additional nutrients:*** This thesis develops the spectral index, the SPI (Soil pH Index) and SOCI (Soil Organic Carbon Index) for two essential soil properties — soil pH and SOC, and validates using hyperspectral data. However, other important macro- and micronutrients also

play a critical role in evaluating soil fertility and crop productivity. Future research should focus on creating nutrient-specific spectral indices for these elements. Given the unique absorption characteristics and interactions with soil moisture and texture, these nutrients may require different wavelength combinations and modeling strategies than those used for pH and SOC.

2. ***Spatio-Temporal soil nutrient modeling:*** Soil nutrients do not vary only with spectral signatures but also change over space and time, influenced by topography, rainfall, land use, and crop rotation. Future studies should consider integrating spectral indices with spatial datasets (e.g., elevation, slope, soil type maps) and temporal data (e.g., crop calendar, temperature, rainfall, and irrigation cycles). It will enable the creation of spatio-temporally aware predictive models, which are more realistic and practical for long-term soil health monitoring and dynamic crop advisories. Incorporating GIS frameworks and temporal analysis will also support scalable decision-making for large geographical areas.

3. ***Use of self-supervised or transfer learning techniques:*** One of the key challenges in applying DL models is the requirement for large labeled datasets, which are often unavailable for many regions. To overcome this, future work can explore self-supervised learning, where the model learns useful spectral representations from unlabeled data and is later fine-tuned on a small amount of labeled data. Similarly, transfer learning can allow pre-trained models to be adapted for local regions with minimal data by reusing learned spectral feature patterns. These approaches will make deep models more accessible and data-efficient.

8.2 Social Impact and Applications

This research transforms soil nutrient analysis from conventional laboratory methods to AI-enabled, spectral-based digital solutions. The proposed models and spectral indices offer significant potential for practical deployment in agriculture, governance, and environmental monitoring. The following subsections highlight key areas where this thesis can have a meaningful impact.

8.2.1 Social Impact

Spectral-AI-based soil nutrient prediction significantly enhances the social impact of agricultural technologies by enabling accessible, non-invasive soil diagnostics, supporting site-specific nutrient management, and promoting sustainable farming practices. These advancements contribute to more equitable, efficient, and environmentally responsible agricultural solutions worldwide, as follows:

1. ***Democratization of soil testing:*** Traditional soil testing methods are expensive, time-consuming, and inaccessible in rural or developing regions. The proposed frameworks eliminate the need for laboratory infrastructure by offering reliable, image-based predictions, making soil diagnostics more inclusive and frequent.

2. ***Improved soil and environmental health:*** Using accurate spectral indices and AI-based models promotes optimized nutrient application, reducing the overuse of fertilizers. It minimizes leaching and runoff, which often contaminate groundwater and degrade ecosystems. As a result, this thesis supports broader environmental goals such as soil conservation and sustainable nutrient cycling.

3. ***Contribution to climate action:*** SOC prediction plays a key role in understanding carbon sequestration potential. Monitoring SOC from hyperspectral imagery allows for climate-smart land use practices and aids in achieving land degradation neutrality and carbon reporting commitments.

4. ***Enabling equity in digital agriculture:*** The tools developed in this thesis help bridge the digital divide in agriculture by removing technological and geographic barriers to soil monitoring. Small and marginal farmers can benefit from soil health assessments without depending on costly or centralized services.

8.2.2 Applications

The models and spectral indices developed in this thesis are well-positioned for practical deployment in various domains of digital agriculture, soil monitoring, environmental assessment, and governance systems. The application spans from the farm level to regional and national planning platforms. The subsequent discussion outlines the major practical applications derived from this thesis.



Figure 8.1: Climate-smart land use for sustainable soil and carbon management.

1. ***Climate-Smart land use and environmental planning:*** Soil Organic Carbon (SOC) directly indicates soil health, productivity, and carbon sequestration. Advanced soil carbon prediction using models and the SOCI index enables planners to monitor soil carbon, track land degradation, and assess the suitability of land for regenerative agriculture. This thesis directly supports several United Nations Sustainable Development Goals, notably **SDG 2: Zero Hunger, SDG 13: Climate Action, and SDG 15: Life on Land** as shown in Figure 8.1. By enabling the accurate, non-invasive, and scalable prediction of soil nutrients, the proposed model, frameworks, and spectral indices promote site-specific nutrient management, enhance crop productivity, and improve food security in resource-limited regions (SDG 2). The ability to monitor Soil Organic Carbon using the SOCI index contributes to climate-smart agriculture and carbon sequestration efforts (SDG 13). Additionally, by reducing over-fertilization and supporting sustainable soil use, this work helps protect soil biodiversity and prevent land degradation, aligning with global land restoration and conservation goals (SDG 15).

2. ***Precision agriculture systems:*** One of the most direct applications of this thesis lies in precision agriculture, where decisions on fertilizer and water



Figure 8.2: AI-driven precision farming using drones and sensors [21].



Figure 8.3: Digital platforms empower data-driven and smart soil nutrient management.

application are made based on localized variability in soil and crop conditions. The prediction model (HyPLS-GPR) and frameworks (CNViS-Net and CAT-MTLNet) developed using multispectral and hyperspectral data enable fine-grained predictions of soil nutrients. Farm-level predictions enable zone-based nutrient application, minimising fertiliser waste. It supports variable-rate input tailored to soil variability, boosting crop yields and reducing environmental impact, especially in over-fertilized or degraded areas. These capabilities can be integrated into on-field decision support systems used by agronomists, service providers, or autonomous machines (e.g., innovative sprayers, drones, or robotic seeders) as shown in Figure 8.2.

3. ***Integration into digital agriculture platforms:*** Digital platforms are becoming increasingly common in agriculture, especially under national programs like the Soil Health Card scheme (India) or private digital farming tools (e.g., DeHaat, BharatAgri) as shown in Figure 8.3. The proposed spectral indices (SPI and



Figure 8.4: UAV-assisted spectral monitoring for soil nutrient prediction.

SOCI), model (HyPLS-GPR), and frameworks (CNViSNet and CAT-MTLNet) are lightweight, interpretable, and compatible with widely used remote sensing data. These can be easily integrated into such platforms to provide automated soil health reports using imagery as input, eliminating the need for lab data. It classifies nutrient status (low, medium, high) for specific locations and provides advisories, such as lime application for acidic soils or composting for SOC-deficient fields. These platforms can also offer historical dashboards to track changes in soil health over time and include farmer training modules that utilize visual maps to explain the impacts on soil health. This integration can also be linked with crop calendars, weather forecasts, and pest/disease models to build holistic agronomic advisory systems.

4. ***Use in UAV- and drone-based survey systems:*** With the increasing use of drone-mounted hyperspectral sensors, the proposed ML model (HyPLS-GPR), DL-based frameworks (CNViS-Net and CAT-MTLNet), and indices (SPI and SOCI) can be deployed for real-time nutrient mapping during UAV-based surveys. It enables on-the-fly nutrient prediction in fields within minutes after flight. Using dynamic spectral analysis, it can actively monitor soil and crop health at different growth stages—before sowing, mid-season, and after harvest. UAVs capture reflectance data in real-time and transmit it to cloud-based systems. These



Figure 8.5: Data-driven soil nutrient models guiding agricultural policy and subsidies.

systems process the data using trained models and instantly generate spatial nutrient prediction maps. Such applications are especially useful in large-scale farming setups and innovative pilot projects, as shown in Figure 8.4.

5. ***Agricultural policy and subsidy planning:*** The proposed model, frameworks, and spectral indices can support policymakers in identifying nutrient-deficient regions, enabling more targeted and efficient distribution of subsidies. By forecasting regional fertilizer needs and monitoring changes in soil health over time, these tools help optimize resource allocation and evaluate the impact of agricultural schemes, leading to more data-driven and equitable policy decisions. Policy planners and agricultural economists can use these approaches to estimate region-specific fertilizer needs, reducing over- or under-allocation. It identifies nutrient deficiencies at the block or district level to support equitable input distribution, and the proposed approaches help assess the effectiveness of soil improvement schemes through long-term analysis of nutrient trends. The outputs can be fed into state-level soil health dashboards or policy frameworks, such as the National Mission on Sustainable Agriculture (NMSA) and Rashtriya Krishi Vikas Yojana (RKVY), as shown in Figure 8.5.

References

[1] Fernandez-Ugalde *et al.*, "Assessment of Changes in Topsoil Properties in Lucas Samples Between 2009/2012 and 2015 Surveys," *JRC, Luxembourg*, 2020.

[2] Jones *et al.*, "Lucas 2015 Topsoil Survey," *Presentation of Dataset and Results, EUR*, vol. 30332, no. 10.2760, p. 616084, 2020.

[3] J. Leenaars, "Africa Soil Profiles Database, Version 1.1. A Compilation of Georeferenced and Standardised Legacy Soil Profile Data for Sub-Saharan Africa (With Dataset). Africa Soil Information Service (AfSIS) Project." ISRIC-World Soil Information, Technical Report, 2013, Last accessed on August 18, 2025. [Online]. Available: https://www.isric.org/projects/africa-soil-profiles-database-afsp

[4] Nalepa *et al.*, "The Hyperview Challenge: Estimating Soil Parameters From Hyperspectral Images," in *IEEE Proc. International Conference on Image Processing (ICIP)*, 2022, pp. 4268–4272.

[5] Tripathi *et al.*, "A Deep Learning Multi-layer Perceptron and Remote Sensing Approach for Soil Health Based Crop Yield Estimation," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 102959, 2022.

[6] Rivadeneira *et al.*, "Database of Hyperspectral Images of Phosphorus in Soil," Mendeley Data, https://data.mendeley.com/datasets/fvgswvt5ws/3, 2022, Last accessed on September 20, 2025.

[7] Jafarbiglu *et al.*, "A Comprehensive Review of Remote Sensing Platforms, Sensors, and Applications in Nut Crops," *Computers and Electronics in Agriculture*, vol. 197, p. 106844, 2022.

[8] Zhang *et al.*, "A Review on Unmanned Aerial Vehicle Remote Sensing: Platforms, Sensors, Data Processing Methods, and Applications," *Drones*, vol. 7, no. 6, p. 398, 2023.

[9] "Hyperspectral vs Multispectral Imaging: What's the Difference?" https://www.prophotonix.com/blog/hyperspectral-vs-multispectral-imaging-whats-the-difference/, Last accessed on September 5, 2023.

[10] A. Singh. (2024) Hyperspectral vs Multispectral Imaging: Advantages and Disadvantages. https://www.azooptics.com/Article.aspx?ArticleID=2590. Last accessed on August 16, 2025.

[11] M. Sharma, "What is Active and Passive Remote Sensing?" https://geospatialworld.net/videos/active-and-passive-remote-sensing/, 2019, Last accessed on September 14, 2025.

[12] Li *et al.*, "Review of Spectral Imaging Technology in Biomedical Engineering: Achievements and Challenges," *Journal of Biomedical Optics*, vol. 18, no. 10, pp. 100 901–100 901, 2013.

[13] Bandopadhyay *et al.*, "Review of Top-of-Canopy Sun-Induced Fluorescence (SIF) Studies From Ground, UAV, Airborne to Spaceborne Observations," *Sensors*, vol. 20, no. 4, p. 1144, 2020.

[14] Tian *et al.*, "Review of Remote Sensing-Based Methods for Forest Aboveground Biomass Estimation: Progress, Challenges, and Prospects," *Forests*, vol. 14, no. 6, p. 1086, 2023.

[15] J. Brodersen, "Multispectral Lighting Solves Difficult Imaging Challenges," https://www.photonics.com/Articles/Multispectral_Lighting_A_Superior_Option_for/a66251, Last accessed on September 14, 2025.

[16] "Multispectral and Hyperspectral Comparison," https://www.edmundoptics.com/knowledge-center/application-notes/imaging/hyperspectral-and-multispectral-imaging/, Last accessed on September 20, 2023.

[17] Zelenka *et al.*, "Ai4eo Hyperview: A Spectralnet3d and Rnnplus Approach for Sustainable Soil Parameter Estimation on Hyperspectral Image Data," in *IEEE Proc. International Conference on Image Processing (ICIP)*, 2022, pp. 4263–4267.

[18] Rivadeneira-Bolaños *et al.*, "Automated Prediction of Phosphorus Concentration in Soils Using Reflectance Spectroscopy and Machine Learning Algorithms," *MethodsX*, vol. 13, p. 102996, 2024.

[19] Kneissl *et al.*, "Map-Projection-Independent Crater Size-Frequency Determination in GIS Environments—New Software Tool for ArcGIS," *Planetary and Space Science*, vol. 59, no. 11-12, pp. 1243–1254, 2011.

[20] "Landsat 8 Bands and Band Combinations," https://gisgeography.com/landsat-8-bands-combinations/, Last accessed on September 20, 2025.

[21] H. Malone, "Advancing Agriculture: The Revolutionary Impact of Precision Farming," https://humphreymalone.com/2024/01/19/uncategorized/advancing-agriculture-the-revolutionary-impact-of-precision-farming/, 2024, Last accessed on October 5, 2025.

[22] Dethier *et al.*, "Agriculture and Development: A Brief Review of the Literature," *Economic Systems*, vol. 36, no. 2, pp. 175–205, 2012.

[23] "United Nations, The Sustainable Development Goals Report," https://unstats.un.org/sdgs/report/2023/The-Sustainable-Development-Goals-Report-2023.pdf, 2023, Last accessed on November 11, 2023.

[24] Fróna *et al.*, "The Challenge of Feeding the World," *Sustainability*, vol. 11, no. 20, p. 5816, 2019.

[25] W. Busscher, "Spending Our Water and Soils for Food Security," *Journal of Soil and Water Conservation*, vol. 67, no. 3, pp. 228–234, 2012.

[26] P. Koyuncuoğlu and G. Erden, "Sampling, Pre-treatment, and Identification Methods of Microplastics in Sewage Sludge and Their Effects in Agricultural

Soils: A Review," *Environmental Monitoring and Assessment*, vol. 193, pp. 1–28, 2021.

[27] Bünemann *et al.*, "Soil Quality–A Critical Review," *Soil Biology and Biochemistry*, vol. 120, pp. 105–125, 2018.

[28] U. B. M. L. and A. Musa Mala, "Impact of Macronutrients and Micronutrients on Soil Health Nutrients Concentration and Uptake by Maize (Zea Mays L)," *International Journal of Current Microbiology and Applied Sciences*, vol. 10(09), pp. 199–206, 2021.

[29] Yang *et al.*, "Soil Nutrient Estimation and Mapping in Farmland Based on UAV Imaging Spectrometry," *Sensors*, vol. 21, no. 11, p. 3919, 2021.

[30] Gulhane *et al.*, "Correlation Analysis of Soil Nutrients and Prediction Model Through Iso Cluster Unsupervised Classification With Multispectral Data," *Multimedia Tools and Applications*, vol. 82, no. 2, pp. 2165–2184, 2023.

[31] Diaz-Gonzalez *et al.*, "Machine Learning and Remote Sensing Techniques Applied to Estimate Soil Indicators–Review," *Ecological Indicators*, vol. 135, p. 108517, 2022.

[32] Ghazali *et al.*, "Generating Soil Salinity, Soil Moisture, Soil pH From Satellite Imagery and Its Analysis," *Information Processing in Agriculture*, vol. 7, no. 2, pp. 294–306, 2020.

[33] Guo *et al.*, "Mapping of Soil pH Based on SVM-RFE Feature Selection Algorithm," *Agronomy*, vol. 12, no. 11, p. 2742, 2022.

[34] Nawar *et al.*, "Machine Learning Based on-line Prediction of Soil Organic Carbon After Removal of Soil Moisture Effect," *Remote Sensing*, vol. 12, no. 8, p. 1308, 2020.

[35] Yang *et al.*, "Evaluation of Machine Learning Approaches to Predict Soil Organic Matter and Ph Using Vis-NIR Spectra," *Sensors*, vol. 19, no. 2, p. 263, 2019.

[36] Yu *et al.*, "Hyperspectral Remote Sensing Applications in Soil: A review," *Hyperspectral Remote Sensing*, pp. 269–291, 2020.

[37] Nakachew *et al.*, "Review on Enhancing the Efficiency of Fertilizer Utilization: Strategies for Optimal Nutrient Management," *Open Agriculture*, vol. 9, no. 1, p. 20220356, 2024.

[38] Maurya *et al.*, "Indicators for Assessment of Soil Quality: A Mini-Review," *Environmental Monitoring and Assessment*, vol. 192, pp. 1–22, 2020.

[39] Chandraprabha *et al.*, "Soil Based Prediction for Crop Yield Using Predictive Analytics," in *IEEE Proc. 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2021, pp. 265–270.

[40] Misbah *et al.*, "Multi-sensors Remote Sensing Applications for Assessing, Monitoring, and Mapping NPK Content in Soil and Crops in African Agricultural Land," *Remote Sensing*, vol. 14, no. 1, p. 81, 2021.

[41] He *et al.*, "Prediction of Soil Macronutrients Content Using Near-Infrared Spectroscopy," *Computers and Electronics in Agriculture*, vol. 58, no. 2, pp. 144–153, 2007.

[42] U. E. Farwa *et al.*, "Prediction of Soil Macronutrients Using Machine Learning Algorithm," *International Journal of Computer (IJC)*, vol. 38, no. 1, pp. 1–14, 2020.

[43] Sunori *et al.*, "Machine Learning Based Prediction of Soil pH," in *IEEE Proc. 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2021, pp. 884–889.

[44] Zadeh *et al.*, "Spatial Prediction of Soil Organic Carbon Using Machine Learning Techniques in Western Iran," *Geoderma Regional*, vol. 21, p. e00260, 2020.

[45] Wilhelm *et al.*, "Predicting Measures of Soil Health Using the Microbiome and Supervised Machine Learning," *Soil Biology and Biochemistry*, vol. 164, p. 108472, 2022.

[46] Pouladi *et al.*, "Digital Mapping of Soil Organic Carbon Using Remote Sensing Data: A Systematic Review," *Catena*, vol. 232, p. 107409, 2023.

[47] Odebiri *et al.*, "Deep Learning Approaches in Remote Sensing of Soil Organic Carbon: A Review of Utility, Challenges, and Prospects," *Environmental Monitoring and Assessment*, vol. 193, pp. 1–18, 2021.

[48] Abdulraheem *et al.*, "Advancement of Remote Sensing for Soil Measurements and Applications: A Comprehensive Review," *Sustainability*, vol. 15, no. 21, p. 15444, 2023.

[49] "Estimation of Soil Moisture at Different Soil Levels Using Machine Learning Techniques and Unmanned Aerial Vehicle (UAV) Multispectral Imagery, author=Aboutalebi and others," in *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping IV*, vol. 11008. SPIE, 2019, pp. 216–226.

[50] Babbar *et al.*, "Spatial Analysis of Soil Quality Using Geospatial Techniques in Botanic Garden of Indian Republic, Noida, Uttar Pradesh, India," *Environmental Sustainability*, vol. 5, no. 4, pp. 471–492, 2022.

[51] Ballabio *et al.*, "Copper Distribution in European Topsoils: An Assessment Based on Lucas Soil Survey," *Science of the Total Environment*, vol. 636, pp. 282–298, 2018.

[52] Dangal *et al.*, "Accurate and Precise Prediction of Soil Properties From a Large Mid-Infrared Spectral Library," *Soil Systems*, vol. 3, no. 1, p. 11, 2019.

[53] Zhang *et al.*, "A Survey on Computational Spectral Reconstruction Methods From RGB to Hyperspectral Imaging," *Scientific Reports*, vol. 12, no. 1, p. 11905, 2022.

[54] Yamaguchi *et al.*, "Beyond Red–Green–Blue (RGB): Spectrum-Based Color Imaging Technology," *Journal of Imaging Science and Technology*, vol. 52, no. 1, pp. 10 201–1, 2008.

[55] Liu *et al.*, "A Review on Soil Nitrogen Sensing Technologies: Challenges, Progress and Perspectives," *Agriculture*, vol. 13, no. 4, p. 743, 2023.

[56] U. Barman and R. D. Choudhury, "Prediction of Soil PH Using Smartphone Based Digital Image Processing and Prediction Algorithm," *Journal of Mechanics of Continua and Mathematical Sciences*, vol. 14, pp. 226–249, 2019.

[57] Natarajan *et al.*, "Prediction of Soil Ph From Remote Sensing Data Using Gradient Boosted Regression Analysis ," *Journal of Pharmaceutical Negative Results*, pp. 29–36, 2022.

[58] Gorthi *et al.*, "Soil Organic Matter Prediction Using Smartphone-Captured Digital Images: Use of Reflectance Image and Image Perturbation," *Biosystems Engineering*, vol. 209, pp. 154–169, 2021.

[59] Fu *et al.*, "Predicting Soil Organic Matter From Cellular Phone Images Under Varying Soil Moisture," *Geoderma*, vol. 361, p. 114020, 2020.

[60] Lavanya *et al.*, "Digital Soil Mapping of Available Phosphorus Using a Smartphone-Integrated RGB Imaging Device and Ascorbic Acid Extraction Method," *Smart Agricultural Technology*, vol. 9, p. 100591, 2024.

[61] Garini *et al.*, "Spectral imaging: Principles and applications," *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, vol. 69, no. 8, pp. 735–747, 2006.

[62] Adityayuda *et al.*, "Development of a Remote Sensing System for Real-Time Detection of Military Threats," in *IEEE Proc. Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS)*, 2024, pp. 257–268.

[63] Kalambukattu *et al.*, "Digital Soil Mapping in a Himalayan Watershed Using Remote Sensing and Terrain Parameters Employing Artificial Neural Network Model," *Environmental Earth Sciences*, vol. 77, pp. 1–14, 2018.

[64] Saberioon *et al.*, "Enhancing Soil Organic Carbon Prediction of Lucas Soil Database Using Deep Learning and Deep Feature Selection," *Computers and Electronics in Agriculture*, vol. 227, p. 109494, 2024.

[65] Zhao *et al.*, "Estimation of Soil Organic Carbon Content Based on Deep Learning and Quantile Regression," in *IEEE Proc. International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3717–3720.

[66] Abdoli *et al.*, "Use of Remote Sensing Data to Predict Soil Organic Carbon in Some Agricultural Soils of Iran," *Remote Sensing Applications: Society and Environment*, vol. 30, p. 100969, 2023.

[67] Yan *et al.*, "High-Resolution Mapping of Soil Organic Matter at the Field Scale Using UAV Hyperspectral Images with a Small Calibration Dataset," *Remote Sensing*, vol. 15, no. 5, p. 1433, 2023.

[68] K. Garajeh *et al.*, "Spatiotemporal Monitoring of Climate Change Impacts on Water Resources Using an Integrated Approach of Remote Sensing and Google Earth Engine," *Scientific Reports*, vol. 14, no. 1, p. 5469, 2024.

[69] Lee *et al.*, "Determination of Primary Spectral Bands for Remote Sensing of Aquatic Environments," *Sensors*, vol. 7, no. 12, pp. 3428–3441, 2007.

[70] He *et al.*, "Soil Organic Carbon Prediction Using Phenological Parameters and Remote Sensing Variables Generated From Sentinel-2 Images," *Catena*, vol. 205, p. 105442, 2021.

[71] S. Aggarwal, "Principles of Remote Sensing," *Satellite Remote Sensing and GIS Applications in Agricultural Meteorology*, vol. 23, no. 2, pp. 23–28, 2004.

[72] M. A. Mulders, *Remote Sensing in Soil Science*. Elsevier, 1987, vol. 15.

[73] R. S. Dwivedi, *Remote Sensing of Soils*. Springer, 2017, vol. 497.

[74] Boerner *et al.*, "User Collision in Sharing of Electromagnetic Spectrum: Frequency Allocation, RF Interference Reduction and RF Security Threat Mitigation in Radio Propagation and Passive and Active Remote Sensing," in *URSI-F Open Symp*, 2002.

[75] Huang *et al.*, "Space-Based Electromagnetic Spectrum Sensing and Situation Awareness," *Space: Science & Technology*, vol. 4, p. 0109, 2024.

[76] J. Norgard and G. L. Best, "The Electromagnetic Spectrum," in *National Association of Broadcasters Engineering Handbook*. Routledge, 2017, pp. 3–10.

[77] I. Scotford and P. Miller, "Applications of Spectral Reflectance Techniques in Northern European Cereal Production: A Review," *Biosystems engineering*, vol. 90, no. 3, pp. 235–250, 2005.

[78] J. C. Price, "How Unique Are Spectral Signatures?" *Remote Sensing of Environment*, vol. 49, no. 3, pp. 181–186, 1994.

[79] Yuvaraj *et al.*, "Performance of Spectral Indices for Soil Properties: A Case Study From Redland Farm, South Florida," *Modeling Earth Systems and Environment*, vol. 8, no. 4, pp. 4829–4841, 2022.

[80] "Introduction to Spatial and Spectral Resolution: Multispectral Imagery," https://www.earthdatascience.org/courses/earth-analytics/multispectral-remote-sensing-data/introduction-multispectral-imagery-r/, Last accessed on July 2, 2025.

[81] Ma *et al.*, "Multispectral Imaging: Review of Current Applications," *Survey of Ophthalmology*, vol. 68, no. 5, pp. 889–904, 2023.

[82] ElMasry *et al.*, "Principles of Hyperspectral Imaging Technology," in *Hyperspectral Imaging for Food Quality Analysis and Control*. Elsevier, 2010, pp. 3–43.

[83] Prasad *et al.*, "Introduction to the Issue on Advances in Hyperspectral Data Processing and Analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 961–963, 2015.

[84] T. W. Randolph, "Scale-Based Normalization of Spectral Data," *Cancer Biomarkers*, vol. 2, no. 3-4, pp. 135–144, 2006.

[85] Balabin *et al.*, "Variable Selection in Near-Infrared Spectroscopy: Benchmarking of Feature Selection Methods on Biodiesel Data," *Analytica Chimica Acta*, vol. 692, no. 1-2, pp. 63–72, 2011.

[86] Chlouveraki *et al.*, "Estimation of Soil Properties Using Hyperspectral Imaging and Machine Learning," *Smart Agricultural Technology*, vol. 10, p. 100790, 2025.

[87] Xu *et al.*, "Optimizing Machine Learning Models for Predicting Soil pH and Total P in Intact Soil Profiles With Visible and Near-Infrared Reflectance(VNIR) Spectroscopy," *Computers and Electronics in Agriculture*, vol. 218, p. 108643, 2024.

[88] Wang *et al.*, "Non-linear Memory-Based Learning for Predicting Soil Properties Using a Regional Vis-NIR Spectral Library," *Geoderma*, vol. 441, p. 116752, 2024.

[89] E. Jamaoui *et al.*, "Machine Learning-Driven Modeling for Soil Organic Carbon Estimation From Multispectral Drone Imaging: A Case Study in Corvera, Murcia (Spain)," *Modeling Earth Systems and Environment*, pp. 1–22, 2024.

[90] Zhang *et al.*, "On Retrieving the Chromium and Zinc Concentrations in the Arable Soil by the Hyperspectral Reflectance Based on the Deep Forest," *Ecological Indicators*, vol. 144, p. 109440, 2022.

[91] Xu *et al.*, "A Comparison of Machine Learning Algorithms for Mapping Soil Iron Parameters Indicative of Pedogenic Processes by Hyperspectral Imaging of Intact Soil Profiles," *European Journal of Soil Science*, vol. 73, no. 1, p. e13204, 2022.

[92] Zhang *et al.*, "Estimating Temporal Changes in Soil pH in the Black Soil Region of Northeast China Using Remote Sensing," *Computers and Electronics in Agriculture*, vol. 154, pp. 204–212, 2018.

[93] Wang *et al.*, "Hyperspectral Monitor of Soil Chromium Contaminant Based on Deep Learning Network Model in the Eastern Junggar Coalfield," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 257, p. 119739, 2021.

[94] D. Jong *et al.*, "Basics of Remote Sensing," in *Remote Sensing Image Analysis: Including the Spatial Domain*. Springer, 2004, no. 5, pp. 1–15.

[95] Andersen *et al.*, "Active Remote Sensing," in *Computer Applications in Sustainable Forest Management: Including Perspectives on Collaboration and Integration*. Springer, 2006, pp. 43–66.

[96] Jia *et al.*, "Review on Active and Passive Remote Sensing Techniques for Road Extraction," *Remote Sensing*, vol. 13, no. 21, p. 4235, 2021.

[97] Legleiter *et al.*, "Passive Optical Remote Sensing of River Channel Morphology and In-Stream Habitat: Physical Basis and Feasibility," *Remote Sensing of Environment*, vol. 93, no. 4, pp. 493–510, 2004.

[98] Navalgund *et al.*, "Remote Sensing Applications: An Overview," *Current Science*, pp. 1747–1766, 2007.

[99] Wang *et al.*, "Remote Sensing Revolutionizing Agriculture: Toward a New Frontier," *Future Generation Computer Systems*, p. 107691, 2025.

[100] Allu *et al.*, "Impact of Remote Sensing Data Fusion on Agriculture Applications: A Review," *European Journal of Agronomy*, vol. 164, p. 127478, 2025.

[101] Martos *et al.*, "Ensuring Agricultural Sustainability Through Remote Sensing in the Era of Agriculture 5.0," *Applied Sciences*, vol. 11, no. 13, p. 5911, 2021.

[102] Fassnacht *et al.*, "Remote Sensing in Forestry: Current Challenges, Considerations and Directions," *Forestry: An International Journal of Forest Research*, vol. 97, no. 1, pp. 11–37, 2024.

[103] Feng *et al.*, "Security of Target Recognition for UAV Forestry Remote Sensing Based on Multi-Source Data Fusion Transformer Framework," *Information Fusion*, vol. 112, p. 102555, 2024.

[104] Aguilar *et al.*, "Forestry applications of space-borne LiDAR sensors: A worldwide bibliometric analysis," *Sensors*, vol. 24, no. 4, p. 1106, 2024.

[105] Lechner *et al.*, "Applications in Remote Sensing to Forest Ecology and Management," *One Earth*, vol. 2, no. 5, pp. 405–412, 2020.

[106] Sigopi *et al.*, "Advancements in Remote Sensing Technologies for Accurate Monitoring and Management of Surface Water Resources in Africa: An Overview, Limitations, and Future Directions," *Geocarto International*, vol. 39, no. 1, p. 2347935, 2024.

[107] Günen *et al.*, "Remote Sensing and Monitoring of Water Resources: A Comparative Study of Different Indices and Thresholding Methods," *Science of the Total Environment*, vol. 926, p. 172117, 2024.

[108] Metwaly *et al.*, "Digital Mapping of Soil Quality and Salt-affected Soil Indicators for Sustainable Agriculture in the Nile Delta Region," *Remote Sensing Applications: Society and Environment*, vol. 36, p. 101318, 2024.

[109] Bonchkovskyi *et al.*, "War-induced Soil Disturbances in North-eastern Ukraine (Kharkiv Region): Physical Disturbances, Soil Contamination and Land Use Change," *Science of The Total Environment*, vol. 964, p. 178594, 2025.

[110] Zhao *et al.*, "Evaluation of Land Use Land Cover Changes in Response to Land Surface Temperature With Satellite Indices and Remote Sensing Data," *Rangeland Ecology & Management*, vol. 96, pp. 183–196, 2024.

[111] Jadhav *et al.*, "Spatiotemporal Data Fusion and Deep Learning for Remote Sensing-based Sustainable Urban Planning," *International Journal of System Assurance Engineering and Management*, pp. 1–9, 2024.

[112] Jia *et al.*, "Semantic Segmentation of Deep Learning Remote Sensing Images Based on Band Combination Principle: Application in Urban Planning and Land Use," *Computer Communications*, vol. 217, pp. 97–106, 2024.

[113] Cetin *et al.*, "Determination of Land Surface Temperature and Urban Heat Island Effects With Remote Sensing Capabilities: The Case of Kayseri, TÜrkiye," *Natural Hazards*, vol. 120, no. 6, pp. 5509–5536, 2024.

[114] Kumari *et al.*, "A Comprehensive Review of Remote Sensing Technologies for Improved Geological Disaster Management," *Geological Journal*, vol. 60, no. 1, pp. 223–235, 2025.

[115] Park *et al.*, "National Disaster Management and Monitoring Using Satellite Remote Sensing and Geo-information," *Korean Journal of Remote Sensing*, vol. 40, no. 5_2, pp. 813–832, 2024.

[116] R. Ahmad, "Smart Remote Sensing Network for Disaster Management: An Overview," *Telecommunication Systems*, vol. 87, no. 1, pp. 213–237, 2024.

[117] Z. Cetin *et al.*, "The Effects of Climate on Land Use/cover: A Case Study in Turkey by Using Remote Sensing Data," *Environmental Science and Pollution Research*, vol. 30, no. 3, pp. 5688–5699, 2023.

[118] Hussain *et al.*, "Spatiotemporal Variation in Land Use Land Cover in the Response to Local Climate Change Using Multispectral Remote Sensing Data," *Land*, vol. 11, no. 5, p. 595, 2022.

[119] Yang *et al.*, "Assessing the Effects of 2d/3D Urban Morphology on the 3D Urban Thermal Environment by Using Multi-Source Remote Sensing Data and UAV Measurements: A Case Study of the Snow-Climate City of Changchun, China," *Journal of Cleaner Production*, vol. 321, p. 128956, 2021.

[120] Chen *et al.*, "Remote Sensing Big Data for Water Environment Monitoring: Current Status, Challenges, and Future Prospects," *Earth's Future*, vol. 10, no. 2, p. e2021EF002289, 2022.

[121] Tang *et al.*, "A Remote Sensing Image Rotation Object Detection Approach for Real-Time Environmental Monitoring," *Sustainable Energy Technologies and Assessments*, vol. 57, p. 103270, 2023.

[122] Asadzadeh *et al.*, "UAV-Based Remote Sensing for the Petroleum Industry and Environmental Monitoring: State-of-the-Art and Perspectives," *Journal of Petroleum Science and Engineering*, vol. 208, p. 109633, 2022.

[123] Nafigin *et al.*, "Geological and Mineralogical Mapping Based on Statistical Methods of Remote Sensing Data Processing of Landsat-8: A Case Study in the Southeastern Transbaikalia, Russia," *Sustainability*, vol. 14, no. 15, p. 9242, 2022.

[124] S. Peyghambari and Y. Zhang, "Hyperspectral Remote Sensing in Lithological Mapping, Mineral Exploration, and Environmental Geology: An Updated Review," *Journal of Applied Remote Sensing*, vol. 15, no. 3, pp. 031 501–031 501, 2021.

[125] T. U. Omali, "Utilization of Remote Sensing and GIS in Geology and Mining," *International Journal of Scientific Research in Multidisciplinary Studies*, vol. 7, no. 4, 2021.

[126] Krishnamoorthy *et al.*, "Integrating GIS-Remote Sensing: A Comprehensive Approach to Predict Oceanographic Health and Coastal Dynamics," *Remote Sensing in Earth Systems Sciences*, vol. 8, no. 1, pp. 200–212, 2025.

[127] Li *et al.*, "Validation of the Ocean Wave Spectrum From the Remote Sensing Data of the Chinese–French Oceanography Satellite," *Remote Sensing*, vol. 15, no. 16, p. 3918, 2023.

[128] Aulicino *et al.*, "Remote Sensing Applications in Satellite Oceanography," in *Measurement for the Sea: Supporting the Marine Environment and the Blue Economy*. Springer, 2022, pp. 181–209.

[129] Al-Mulla *et al.*, "Remote Sensing Techniques in the Fields of Defense and Security Studies," *Journal of Strategic and Defense Studies (JSDS)*, vol. 1, no. 1, 2025.

[130] Hutsul *et al.*, "Review of Approaches to the Use of Unmanned Aerial Vehicles, Remote Sensing and Geographic Information Systems in Humanitarian Demining: Ukrainian Case," *Heliyon*, vol. 10, no. 7, 2024.

[131] Odebiri *et al.*, "Basic and Deep Learning Models in Remote Sensing of Soil Organic Carbon Estimation: A Brief Review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102389, 2021.

[132] Wei *et al.*, "Dimensionality Reduction Statistical Models for Soil Attribute Prediction Based on Raw Spectral Data," *AI*, vol. 3, no. 4, pp. 809–819, 2022.

[133] Peng *et al.*, "Prediction of Soil Nutrient Contents Using Visible and Near-infrared Reflectance Spectroscopy," *ISPRS International Journal of Geo-Information*, vol. 8, no. 10, p. 437, 2019.

[134] Gruszczyński *et al.*, "Assessing the Information Potential of MIR Spectral Signatures for Prediction of Multiple Soil Properties Based on Data from the AfSIS Phase I Project," *International Journal of Environmental Research and Public Health*, vol. 19, no. 22, p. 15210, 2022.

[135] Chinilin *et al.*, "Vis-NIR Spectroscopy for Soil Organic Carbon Assessment: A Meta-Analysis," *Eurasian Soil Science*, vol. 56, no. 11, pp. 1605–1617, 2023.

[136] Bellon-Maurel *et al.*, "Near-Infrared (NIR) and Mid-Infrared (MIR) Spectroscopic Techniques for Assessing the Amount of Carbon Stock in Soils–Critical Review and Research Perspectives," *Soil Biology and Biochemistry*, vol. 43, no. 7, pp. 1398–1410, 2011.

[137] Asadzadeh *et al.*, "A Review on Spectral Processing Methods for Geological Remote Sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 47, pp. 69–90, 2016.

[138] Liu *et al.*, "Applications of Spectral Reflectance Techniques in Northern European Cereal Production: A Review," *Chinese Journal of Plant Ecology*, vol. 40, no. 1, p. 80, 2016.

[139] Teillet *et al.*, "Effects of Spectral, Spatial, and Radiometric Characteristics on Remote Sensing Vegetation Indices of Forested Regions," *Remote Sensing of Environment*, vol. 61, no. 1, pp. 139–149, 1997.

[140] Bioucas-Dias *et al.*, "Hyperspectral Remote Sensing Data Analysis and Future Challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[141] Toth *et al.*, "Remote Sensing Platforms and Sensors: A Survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 22–36, 2016.

[142] Guanter *et al.*, "Spectral Calibration and Atmospheric Correction of Ultra-Fine Spectral and Spatial Resolution Remote Sensing Data. Application to CASI-1500 Data," *Remote Sensing of Environment*, vol. 109, no. 1, pp. 54–65, 2007.

[143] Zhang *et al.*, "Dark Background Correction With Multivariate Model Fitting for Infrared Hyperspectral Remote Sensing," *Measurement*, vol. 237, p. 115191, 2024.

[144] Antony *et al.*, "High-Resolution Ultra-Spectral Imager for Advanced Imaging in Agriculture and Biomedical Applications," *Journal of Biomedical Photonics & Engineering*, vol. 9, no. 3, p. 030304, 2023.

[145] Levenson *et al.*, "Multispectral Imaging in Biology and Medicine: Slices of Life," *Cytometry Part A: the journal of the International Society for Analytical Cytology*, vol. 69, no. 8, pp. 748–758, 2006.

[146] Khan *et al.*, "Modern Trends in Hyperspectral Image Analysis: A Review," *IEEE Access*, vol. 6, pp. 14 118–14 129, 2018.

[147] Xu *et al.*, "Hyperspectral Anomaly Detection Based on Machine Learning: An Overview," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 3351–3364, 2022.

[148] Jin *et al.*, "Comparison of Different Satellite Bands and Vegetation Indices for Estimation of Soil Organic Matter Based on Simulated Spectral Configuration," *Agricultural and Forest Meteorology*, vol. 244, pp. 57–71, 2017.

[149] Zhang *et al.*, "Spectral Features Extraction for Estimation of Soil Total Nitrogen Content Based on Modified Ant Colony Optimization Algorithm," *Geoderma*, vol. 333, pp. 23–34, 2019.

[150] Angelopoulou *et al.*, "Remote Sensing Techniques for Soil Organic Carbon Estimation: A Review," *Remote Sensing*, vol. 11, no. 6, p. 676, 2019.

[151] Chu *et al.*, "Artificial Intelligence for Spectral Analysis: Challenges and Opportunities," in *IEEE Proc. International Conference on Big Data (BigData)*, 2023, pp. 5176–5180.

[152] Salimi *et al.*, "Using a Feature Subset Selection Method and Support Vector Machine to Address Curse of Dimensionality and Redundancy in Hyperion Hyperspectral Data Classification," *The Egyptian Journal of Remote Sensing and Space Science*, vol. 21, no. 1, pp. 27–36, 2018.

[153] Shi *et al.*, "Improving Soil Organic Matter Estimation Accuracy by Combining Optimal Spectral Preprocessing and Feature Selection Methods Based on Pxrf and Vis-NIR Data Fusion," *Geoderma*, vol. 430, p. 116301, 2023.

[154] C. V. G. Zelaya, "Towards Explaining the Effects of Data Preprocessing on Machine Learning," in *IEEE Proc. 35th International Conference on Data Engineering (ICDE)*, 2019, pp. 2086–2090.

[155] Zeraatpisheh *et al.*, "Improving the Spatial Prediction of Soil Organic Carbon Using Environmental Covariates Selection: A Comparison of a Group of Environmental Covariates," *Catena*, vol. 208, p. 105723, 2022.

[156] Sharma *et al.*, "Remote Sensing and GIS in Natural Resource Management: Comparing Tools and Emphasizing the Importance of In-Situ Data," *Remote Sensing*, vol. 16, no. 22, p. 4161, 2024.

[157] Moyroud *et al.*, "Introduction to QGIS," *QGIS and Generic Tools*, vol. 1, pp. 1–17, 2018.

[158] M. M. Kooshki, "Introduction to ENVI Software," *The Scientific-Cultural Journal of Ecosphere*, vol. 7, no. 1, pp. 56–59, 2024.

[159] Acharya *et al.*, "Exploring Landsat 8," *International Journal of IT, Engineering and Applied Sciences Research (IJIEASR)*, vol. 4, no. 4, pp. 4–10, 2015.

[160] Ismail *et al.*, "Comparative Assessment of the Unsupervised Land Use Classification by Using Proprietary GIS and Open Source Software," in *IOP Proc. Earth and Environmental Science*, vol. 540, no. 1, 2020, p. 012020.

[161] Meng *et al.*, "Soil Organic Matter Prediction Model With Satellite Hyperspectral Image Based on Optimized Denoising Method," *Remote Sensing*, vol. 13, no. 12, p. 2273, 2021.

[162] Zimmermann *et al.*, "Optimizing Savitzky–Golay Parameters for Improving Spectral Resolution and Quantification in Infrared Spectroscopy," *Applied Spectroscopy*, vol. 67, no. 8, pp. 892–902, 2013.

[163] J. Zhang and A. M. Mouazen, "A Novel Non-Integer Order Savitzky–Golay Derivative Function of Visible and Near-Infrared Spectra for Improving Prediction Accuracy of Phosphorus in Pig Manure," *Chemometrics and Intelligent Laboratory Systems*, vol. 240, p. 104917, 2023.

[164] J. Zhang and A. M. Mouazen, "Fractional-Order Savitzky–Golay Filter for Pre-Treatment of On-Line Vis–NIR Spectra To Predict Phosphorus in Soil," *Infrared Physics & Technology*, vol. 131, p. 104720, 2023.

[165] Jiao *et al.*, "Estimation of Surface Soil Nutrient Content in Mountainous Citrus Orchards Based on Hyperspectral Data," *Agriculture*, vol. 14, no. 6, p. 873, 2024.

[166] Wang *et al.*, "Fourier-Transform Infrared Spectral Inversion of Soil Available Potassium Content Based on Different Dimensionality Reduction Algorithms," *Agronomy*, vol. 13, no. 3, p. 617, 2023.

[167] C. Cobas, "Applications of the Whittaker Smoother in NMR Spectroscopy," *Magnetic Resonance in Chemistry*, vol. 56, no. 12, pp. 1140–1148, 2018.

[168] Zhua *et al.*, "Establishing a Periodic SM Model With Fourier Analysis for Enhancing Global Soil Moisture Forecasting," *Scientific Reports*, vol. 15, no. 1, p. 15962, 2025.

[169] Shao *et al.*, "An Evaluation of Time-Series Smoothing Algorithms for Land-Cover Classifications Using MODIS-NDVI Multi-Temporal Data," *Remote Sensing of Environment*, vol. 174, pp. 258–265, 2016.

[170] Liang *et al.*, "Using Enhanced Gap-Filling and Whittaker Smoothing To Reconstruct High Spatiotemporal Resolution NDVI Time Series Based on Landsat 8, Sentinel-2, and MODIS Imagery," *ISPRS International Journal of Geo-Information*, vol. 12, no. 6, p. 214, 2023.

[171] R. Anand, "Quantum-Enhanced Soil Nutrient Estimation Exploiting Hyperspectral Data with Quantum Fourier Transform," *IEEE Geoscience and Remote Sensing Letters*, 2025.

[172] Krylov *et al.*, "Recovering Fluorescence Spectra Hidden by Scattering Signal: In Search of the Best Smoother," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 293, p. 122441, 2023.

[173] M. A. Siddiqi and W. Pak, "An Agile Approach to Identify Single and Hybrid Normalization for Enhancing Machine Learning-Based Network Intrusion Detection," *IEEE Access*, vol. 9, pp. 137 494–137 513, 2021.

[174] Bi *et al.*, "A Local Pre-processing Method for Near-infrared Spectra, Combined With Spectral Segmentation and Standard Normal Variate Transformation," *Analytica Chimica Acta*, vol. 909, pp. 30–40, 2016.

[175] Shin *et al.*, "A Preprocessing Technique Using Diffuse Reflectance Spectroscopy to Predict the Soil Properties of Paddy Fields in Korea," *Applied Sciences*, vol. 14, no. 11, p. 4673, 2024.

[176] Canero *et al.*, "Machine Learning and Feature Selection for soil spectroscopy. An evaluation of Random Forest wrappers to predict soil organic matter, clay, and carbonates," *Heliyon*, vol. 10, no. 9, 2024.

[177] Vestergaard *et al.*, "Evaluation of Optimized Preprocessing and Modeling Algorithms for Prediction of Soil Properties Using Vis-NIR Spectroscopy," *Sensors*, vol. 21, no. 20, p. 6745, 2021.

[178] Zhang *et al.*, "Automated Pipeline for Classifying Aroclors in Soil by Gas Chromatography/Mass Spectrometry Using Modulo Compressed Two-Way Data Objects," *Talanta*, vol. 117, pp. 483–491, 2013.

[179] R. Popescu, *Stochastic Variability of Soil Properties: Data Analysis, Digital Simulation, Effects on System Behavior*. Princeton University, 1995.

[180] Yemefack *et al.*, "Developing a Minimum Data Set for Characterizing Soil Dynamics in Shifting Cultivation Systems," *Soil and Tillage Research*, vol. 86, no. 1, pp. 84–98, 2006.

[181] Kasemi *et al.*, "EdgeSoil 2.0–Soil Analyzer Using Convolutional Neural Network and Camera Imaging for Agricultural Robotics," in *IEEE Proc. International Conference on Robotics and Automation (ICRA)*, 2024, pp. 15 825–15 831.

[182] Lanjewar *et al.*, "Machine Learning Approach With Data Normalization Technique for Early Stage Detection of Hypothyroidism," in *Artificial Intelligence Applications for Health Care*. CRC Press, 2022, pp. 91–108.

[183] Deng *et al.*, "Union With Recursive Feature Elimination: a Feature Selection Framework to Improve the Classification Performance of Multicategory Causes of Death in Colorectal Cancer," *Laboratory Investigation*, vol. 104, no. 3, p. 100320, 2024.

[184] Chaleshtori *et al.*, "A Novel Bearing Fault Diagnosis Approach Using the Gaussian Mixture Model and the Weighted Principal Component Analysis," *Reliability Engineering & System Safety*, vol. 242, p. 109720, 2024.

[185] Guo *et al.*, "Evaluating Calibration and Spectral Variable Selection Methods for Predicting Three Soil Nutrients Using Vis-NIR Spectroscopy," *Remote Sensing*, vol. 13, no. 19, p. 4000, 2021.

[186] Kirešová *et al.*, "Utilizing Low-cost Optical Sensor for the Measurement of Particulate Matter and Calculating Pearson's Correlation Coefficient," in *SPIE Proc. Sixteenth International Conference on Correlation Optics*, vol. 12938, 2024, pp. 238–246.

[187] Zhang *et al.*, "Estimation of the Total Soil Nitrogen Based on a Differential Evolution Algorithm From ZY1-02D Hyperspectral Satellite Imagery," *Agronomy*, vol. 13, no. 7, p. 1842, 2023.

[188] Sun *et al.*, "Splitting Tensile Strength of Basalt Fiber Reinforced Coral Aggregate Concrete: Optimized Xgboost Models and Experimental Validation," *Construction and Building Materials*, vol. 416, p. 135133, 2024.

[189] Zhao *et al.*, "Estimation of Soil Properties Content with Vis-NIR Spectra by Multi-Task Deep Learning Based on Attention Mechanism and Loss-weight Balancing," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[190] Zhao *et al.*, "Attention-Based CNN Ensemble for Soil Organic Carbon Content Estimation With Spectral Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[191] Zhong *et al.*, "Soil Properties: Their Prediction and Feature Extraction From the Lucas Spectral Library Using Deep Convolutional Neural Networks," *Geoderma*, vol. 402, p. 115366, 2021.

[192] Gacek *et al.*, "Estimating Soil Parameters from Hyperspectral Images using Ensembles of Classic and Deep Machine Learning Models," in *IEEE Proc. International Geoscience and Remote Sensing Symposium IGARSS*, 2024, pp. 412–416.

[193] Wolfrum *et al.*, "A Performance Comparison of Low-Cost Near-Infrared (NIR) Spectrometers to a Conventional Laboratory Spectrometer for Rapid Biomass Compositional Analysis," *BioEnergy Research*, vol. 13, no. 4, pp. 1121–1129, 2020.

[194] JCstat *et al.*, "Africa soil property prediction challenge," https://kaggle.com/competitions/afsis-soil-properties, 2014, kaggle, Last accessed on September, 2025.

[195] Kuzu *et al.*, "Predicting Soil Properties From Hyperspectral Satellite Images," in *IEEE Proc. International Conference on Image Processing (ICIP)*, 2022, pp. 4296–4300.

[196] Wijata *et al.*, "Intuition-1: Toward In-Orbit Bare Soil Detection Using Spectral Vegetation Indices," in *IEEE Proc. International Geoscience and Remote Sensing Symposium IGARSS*, 2024, pp. 1708–1712.

[197] Gholizadeh *et al.*, "Soil Organic Carbon and Texture Retrieving and Mapping Using Proximal, Airborne and Sentinel-2 Spectral Imaging," *Remote Sensing of Environment*, vol. 218, pp. 89–103, 2018.

[198] Dhiman *et al.*, "Soil Textures and Nutrients Estimation Using Remote Sensing Data in North India-Punjab Region," *Procedia Computer Science*, vol. 218, pp. 2041–2048, 2023.

[199] Piccoli *et al.*, "A Deep Scalable Neural Architecture for Soil Properties Estimation From Spectral Information," *Computers & Geosciences*, p. 105433, 2023.

[200] Riad *et al.*, "Prediction of Soil Nutrients Using Hyperspectral Satellite Imaging," in *Springer Proc. International Conference on Fourth Industrial Revolution and Beyond*, 2022, pp. 183–198.

[201] Castaldi *et al.*, "Soil Organic Carbon Mapping Using Lucas Topsoil Database and Sentinel-2 Data: An Approach to Reduce Soil Moisture and Crop Residue Effects," *Remote Sensing*, vol. 11, no. 18, p. 2121, 2019.

[202] S. A. Medjahed and M. Ouali, "Band Selection Based on Optimization Approach for Hyperspectral Image Classification," *The Egyptian Journal of Remote Sensing and Space Science*, vol. 21, no. 3, pp. 413–418, 2018.

[203] Peng *et al.*, "Mapping Vertical Distribution of Soc and TN in Reclaimed Mine Soils Using Point and Imaging Spectroscopy," *Ecological Indicators*, vol. 158, p. 111437, 2024.

[204] Yuzugullu *et al.*, "Satellite-Based Soil Organic Carbon Mapping on European Soils Using Available Datasets and Support Sampling," *Science of Remote Sensing*, vol. 9, p. 100118, 2024.

[205] Guo *et al.*, "Mapping Soil Organic Matter Content Based on Feature Band Selection with ZY1-02D Hyperspectral Satellite Data in the Agricultural Region," *Agronomy*, vol. 12, no. 9, p. 2111, 2022.

[206] Pechanec *et al.*, "Usage of Airborne Hyperspectral Imaging Data for Identifying Spatial Variability of Soil Nitrogen Content," *ISPRS International Journal of Geo-Information*, vol. 10, no. 6, p. 355, 2021.

[207] Dong *et al.*, "Combination of Machine Learning and VIRS for Predicting Soil Organic Matter," *Journal of Soils and Sediments*, vol. 21, no. 7, pp. 2578–2588, 2021.

[208] Zhang *et al.*, "Strategies for the Efficient Estimation of Soil Organic Matter in Salt-Affected Soils Through Vis-NIR Spectroscopy: Optimal Band Combination Algorithm and Spectral Degradation," *Geoderma*, vol. 382, p. 114729, 2021.

[209] Sabetizade *et al.*, "Combination of Mir Spectroscopy and Environmental Covariates to Predict Soil Organic Carbon in a Semi-arid Region," *Catena*, vol. 196, p. 104844, 2021.

[210] Zhu *et al.*, "Assessing Soil Organic Matter Content in a Coal Mining Area Through Spectral Variables of Different Numbers of Dimensions," *Sensors*, vol. 20, no. 6, p. 1795, 2020.

[211] Bangelesa *et al.*, "Predicting Soil Organic Carbon Content Using Hyperspectral Remote Sensing in a Degraded Mountain Landscape in Lesotho," *Applied and Environmental Soil Science*, vol. 2020, pp. 1–11, 2020.

[212] Peng *et al.*, "Estimating Soil Organic Carbon Using VIS/NIR Spectroscopy With SVMR and SPA Methods," *Remote Sensing*, vol. 6, no. 4, pp. 2699–2717, 2014.

[213] Guo *et al.*, "Mapping Soil Organic Matter Content Based on Feature Band Selection With Zy1-02d Hyperspectral Satellite Data in the Agricultural Region," *Agronomy*, vol. 12, no. 9, p. 2111, 2022.

[214] Song *et al.*, "Feature Selection Using Principal Component Analysis," in *IEEE Proc. International Conference on System Science, Engineering Design and Manufacturing Informatization*, vol. 1, 2010, pp. 27–30.

[215] Xie *et al.*, "Soil Type Recognition As Improved by Genetic Algorithm-Based Variable Selection Using Near Infrared Spectroscopy and Partial Least Squares Discriminant Analysis," *Scientific Reports*, vol. 5, no. 1, p. 10930, 2015.

[216] Shekofteh *et al.*, "Optimal Feature Selection for Predicting Soil CEC: Comparing the Hybrid of Ant Colony Organization Algorithm and Adaptive Network-Based

Fuzzy System With Multiple Linear Regression," *Geoderma*, vol. 298, pp. 27–34, 2017.

[217] Yang *et al.*, "Improving the Accuracy of Soil Organic Carbon Estimation: CWT-Random Frog-XGBoost as a Prerequisite Technique for In Situ Hyperspectral Analysis," *Remote Sensing*, vol. 15, no. 22, p. 5294, 2023.

[218] Li *et al.*, "A Fusion XGBoost Approach for Large-Scale Monitoring of Soil Heavy Metal in Farmland Using Hyperspectral Imagery," *Agronomy*, vol. 15, no. 3, p. 676, 2025.

[219] Zhang *et al.*, "Prediction of Soil Organic Matter in Northwestern China Using Fractional-Order Derivative Spectroscopy and Modified Normalized Difference Indices," *Catena*, vol. 185, p. 104257, 2020.

[220] Xie *et al.*, "A Review of Recent Advances for the Detection of Biological, Chemical and Physical Hazards in Foodstuffs Using Spectral Imaging Techniques," *Foods*, vol. 12, no. 11, p. 2266, 2023.

[221] Pellikka *et al.*, "Tropical Altitudinal Gradient Soil Organic Carbon and Nitrogen Estimation Using Specim IQ Portable Imaging Spectrometer," *Science of the Total Environment*, vol. 883, p. 163677, 2023.

[222] Singha *et al.*, "Prediction of Soil Nutrients Through PLSR and SVMR Models by Vis-NIR Reflectance Spectroscopy," *The Egyptian Journal of Remote Sensing and Space Sciences*, vol. 26, no. 4, pp. 901–918, 2023.

[223] Zeraatpisheh *et al.*, "Soil Organic Carbon Stock Prediction Using Multi-Spatial Resolutions of Environmental Variables: How Well Does the Prediction Match Local References?" *Catena*, vol. 229, p. 107197, 2023.

[224] Monsalve *et al.*, "Use of Machine Learning Models for Prediction of Organic Carbon and Nitrogen in Soil from Hyperspectral Imagery in Laboratory," *Journal of Spectroscopy*, vol. 2023, 2023.

[225] Angelopoulou *et al.*, "Evaluation of Airborne Hyspex and Spaceborne PRSIMA Hyperspectral Remote Sensing Data for Soil Organic Matter and Carbonates Estimation," *Remote Sensing*, vol. 15, no. 4, p. 1106, 2023.

[226] Wu *et al.*, "Mapping Soil Organic Carbon in Floodplain Farmland: Implications of Effective Range of Environmental Variables," *Land*, vol. 12, no. 6, p. 1198, 2023.

[227] Luo *et al.*, "Mapping Soil Organic Matter Content Using Sentinel-2 Synthetic Images at Different Time Intervals in Northeast China," *International Journal of Digital Earth*, vol. 16, no. 1, pp. 1094–1107, 2023.

[228] Zhang *et al.*, "Prediction of Soil Organic Carbon Content Using Sentinel-1/2 and Machine Learning Algorithms in Swamp Wetlands in Northeast China," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[229] Chang *et al.*, "Soil Organic Carbon Prediction Based on Different Combinations of Hyperspectral Feature Selection and Regression Algorithms," *Agronomy*, vol. 13, no. 7, p. 1806, 2023.

[230] Liu *et al.*, "Interaction of Climate, Topography and Soil Properties With Cropland and Cropping Pattern Using Remote Sensing Data and Machine Learning Methods," *The Egyptian Journal of Remote Sensing and Space Science*, vol. 26, no. 3, pp. 415–426, 2023.

[231] L. Wang and R. Wang, "Determination of Soil PH From Vis-NIR Spectroscopy by Extreme Learning Machine and Variable Selection: A Case Study in Lime Concretion Black Soil," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 283, p. 121707, 2022.

[232] Wang *et al.*, "Using Soil Library Hyperspectral Reflectance and Machine Learning to Predict Soil Organic Carbon: Assessing Potential of Airborne and Spaceborne Optical Soil Sensing," *Remote Sensing of Environment*, vol. 271, p. 112914, 2022.

[233] Garosi *et al.*, "Effects of Different Sources and Spatial Resolutions of Environmental Covariates on Predicting Soil Organic Carbon Using Machine Learning in a Semi-arid Region of Iran," *Geoderma Regional*, vol. 29, p. e00513, 2022.

[234] Lu *et al.*, "Precise Prediction of Soil Organic Matter in Soils Planted With a Variety of Crops Through Hybrid Methods," *Computers and Electronics in Agriculture*, vol. 200, p. 107246, 2022.

[235] Heil *et al.*, "Fine-Scale Mapping of Soil Organic Matter in Agricultural Soils Using UAVs and Machine Learning," *Remote Sensing*, vol. 14, no. 14, p. 3349, 2022.

[236] Luo *et al.*, "Regional Mapping of Soil Organic Matter Content Using Multitemporal Synthetic Landsat 8 Images in Google Earth Engine," *Catena*, vol. 209, p. 105842, 2022.

[237] Shafizadeh-Moghadam *et al.*, "Synergetic Use of Multi-Temporal Sentinel-1, Sentinel-2, NDVI, and Topographic Factors for Estimating Soil Organic Carbon," *Catena*, vol. 212, p. 106077, 2022.

[238] Phiri *et al.*, "Sentinel-2 Data for Land Cover/use Mapping: A Review," *Remote Sensing*, vol. 12, no. 14, p. 2291, 2020.

[239] Febereiro *et al.*, "Evaluation of Spectroscopy and Methodological Pre-treatments to Estimate Soil Nutrients in the Vineyard," *Remote Sensing*, vol. 14, no. 6, p. 1326, 2022.

[240] Clingensmith *et al.*, "Predicting Soil Properties and Interpreting Vis-NIR Models From Across Continental United States," *Sensors*, vol. 22, no. 9, p. 3187, 2022.

[241] Xie *et al.*, "Estimation of Soil Organic Carbon Content in the Ebinur Lake Wetland, Xinjiang, China, Based on Multisource Remote Sensing Data and Ensemble Learning Algorithms," *Sensors*, vol. 22, no. 7, p. 2685, 2022.

[242] B. others, "Vis-NIR Spectra Combined With Machine Learning for Predicting Soil Nutrients in Cropland From Aceh Province, Indonesia," *Case Studies in Chemical and Environmental Engineering*, vol. 6, p. 100268, 2022.

[243] Wang *et al.*, "Application of Hyperspectral Technology Combined With Bat Algorithm-Adaboost Model in Field Soil Nutrient Prediction," *IEEE Access*, vol. 10, pp. 100 286–100 299, 2022.

[244] Levi *et al.*, "Airborne Imaging Spectroscopy for Assessing Land-use Effect on Soil Quality in Drylands," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 186, pp. 34–54, 2022.

[245] Gasmi *et al.*, "Using PRISMA Hyperspectral Satellite Imagery and GIS Approaches for Soil Fertility Mapping (FertiMap) in Northern Morocco," *Remote Sensing*, vol. 14, no. 16, p. 4080, 2022.

[246] Zhou *et al.*, "Prediction of Total Nitrogen Distribution in Surface Soil Based on Multi-source Auxiliary Variables and Random Forest Approach," Copernicus Meetings, Technical Report, 2021.

[247] Sankey *et al.*, "Quantifying Plant-soil-nutrient Dynamics in Rangelands: Fusion of UAV Hyperspectral-LiDAR, UAV Multispectral-photogrammetry, and Ground-based LiDAR-digital Photography in a Shrub-encroached Desert Grassland," *Remote Sensing of Environment*, vol. 253, p. 112223, 2021.

[248] Yu *et al.*, "Evaluation of Soil Properties, Topographic Metrics, Plant Height, and Unmanned Aerial Vehicle Multispectral Imagery Using Machine Learning Methods to Estimate Canopy Nitrogen Weight in Corn," *Remote Sensing*, vol. 13, no. 16, p. 3105, 2021.

[249] Wang *et al.*, "Assessment of Soil Fertility Degradation Affected by Mining Disturbance and Land Use in a Coalfield via Machine Learning," *Ecological Indicators*, vol. 125, p. 107608, 2021.

[250] Bao *et al.*, "Soil Nutrient Estimation and Mapping in Agriculture Land Based on Improved ELM and UAV Imaging Spectrometry," Copernicus Meetings, Technical Report, 2021.

[251] Xu *et al.*, "Integrating Hyperspectral Imaging With Machine Learning Techniques for the High-Resolution Mapping of Soil Nitrogen Fractions in Soil Profiles," *Science of The Total Environment*, vol. 754, p. 142135, 2021.

[252] Hossen *et al.*, "Total Nitrogen Estimation in Agricultural Soils via Aerial Multispectral Imaging And LIBS," *Scientific Reports*, vol. 11, no. 1, p. 12693, 2021.

[253] Bao *et al.*, "Characterization and Prediction of Soil Organic Matter Content in Reclaimed Mine Soil Using Visible and Near-Infrared Diffuse Spectroscopy," *Arid Land Research and Management*, vol. 35, no. 3, pp. 276–291, 2021.

[254] Zhang *et al.*, "Mapping Regional Soil Organic Matter Based on Sentinel-2A and Modis Imagery Using Machine Learning Algorithms and Google Earth Engine," *Remote Sensing*, vol. 13, no. 15, p. 2934, 2021.

[255] Zhou *et al.*, "A New Coupled Elimination Method of Soil Moisture and Particle Size Interferences on Predicting Soil Total Nitrogen Concentration Through Discrete Nir Spectral Band Data," *Remote Sensing*, vol. 13, no. 4, p. 762, 2021.

[256] N. L. Tsakiridis *et al.*, "Improving the Predictions of Soil Properties From VNIR–SWIR Spectra in an Unlabeled Region Using Semi-supervised and Active Learning," *Geoderma*, vol. 387, p. 114830, 2021.

[257] Wang *et al.*, "Prediction of Soil Organic Carbon Under Different Land Use Types Using Sentinel-1/-2 Data in a Small Watershed," *Remote Sensing*, vol. 13, no. 7, p. 1229, 2021.

[258] Shi *et al.*, "Hyperspectral Band Selection and Modeling of Soil Organic Matter Content in a Forest Using the Ranger Algorithm," *PloS One*, vol. 16, no. 6, p. e0253385, 2021.

[259] Wang *et al.*, "Estimating Soil Organic Matter Content Using Sentinel-2 Imagery by Machine Learning in Shanghai," *IEEE Access*, vol. 9, pp. 78 215–78 225, 2021.

[260] Xu *et al.*, "Improving the Accuracy of Soil Organic Carbon Content Prediction Based on Visible and Near-Infrared Spectroscopy and Machine Learning," *Environmental Earth Sciences*, vol. 80, no. 8, p. 326, 2021.

[261] Lin *et al.*, "Estimation of Soil Total Nitrogen Using the Synthetic Color Learning Machine (SCLM) Method and Hyperspectral Data," *Geoderma*, vol. 380, p. 114664, 2020.

[262] Hong *et al.*, "Exploring the Potential of Airborne Hyperspectral Image for Estimating Topsoil Organic Carbon: Effects of Fractional-Order Derivative and Optimal Band Combination Algorithm," *Geoderma*, vol. 365, p. 114228, 2020.

[263] Zhou *et al.*, "High-Resolution Digital Mapping of Soil Organic Carbon and Soil Total Nitrogen Using Dem Derivatives, Sentinel-1 and Sentinel-2 Data Based on Machine Learning Algorithms," *Science of The Total Environment*, vol. 729, p. 138244, 2020.

[264] Zhou *et al.*, "Mapping Soil Organic Carbon Content Using Multi-Source Remote Sensing Variables in the Heihe River Basin in China," *Ecological Indicators*, vol. 114, p. 106288, 2020.

[265] Wang *et al.*, "Predicting Soil Organic Carbon Content in Spain by Combining Landsat TM and ALOS PALSAR Images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 92, p. 102182, 2020.

[266] Ballabio *et al.*, "Mapping LUCAS Topsoil Chemical Properties at European Scale Using Gaussian Process Regression," *Geoderma*, vol. 355, p. 113912, 2019.

[267] Hong *et al.*, "Estimating Lead and Zinc Concentrations in Peri-Urban Agricultural Soils Through Reflectance Spectroscopy: Effects of Fractional-Order Derivative and Random Forest," *Science of the Total Environment*, vol. 651, pp. 1969–1982, 2019.

[268] Dou *et al.*, "Prediction of Soil Organic Matter Using Multi-Temporal Satellite Images in the Songnen Plain, China," *Geoderma*, vol. 356, p. 113896, 2019.

[269] Li *et al.*, "Quantitative Analysis of Soil Total Nitrogen Using Hyperspectral Imaging Technology With Extreme Learning Machine," *Sensors*, vol. 19, no. 20, p. 4355, 2019.

[270] Moura-Bueno *et al.*, "Stratification of a Local Vis-NIR-SWIR Spectral Library by Homogeneity Criteria Yields More Accurate Soil Organic Carbon Predictions," *Geoderma*, vol. 337, pp. 565–581, 2019.

[271] Castaldi *et al.*, "Estimation of Soil Organic Carbon in Arable Soil in Belgium and Luxembourg With the Lucas Topsoil Database," *European Journal of Soil Science*, vol. 69, no. 4, pp. 592–603, 2018.

[272] Song *et al.*, "Predicting Spatial Variations in Soil Nutrients With Hyperspectral Remote Sensing at Regional Scale," *Sensors*, vol. 18, no. 9, p. 3086, 2018.

[273] Hong *et al.*, "Rapid Identification of Soil Organic Matter Level via Visible and Near-Infrared Spectroscopy: Effects of Two-Dimensional Correlation Coefficient and Extreme Learning Machine," *Science of the Total Environment*, vol. 644, pp. 1232–1243, 2018.

[274] Yue *et al.*, "Black Soil Organic Matter Content Estimation Using Hybrid Selection Method Based on Rf and Gabpso," *Spectroscopy and Spectral Analysis*, vol. 38, no. 1, pp. 181–187, 2018.

[275] Xu *et al.*, "Comparison of Multivariate Methods for Estimating Selected Soil Properties From Intact Soil Cores of Paddy Fields by Vis–NIR Spectroscopy," *Geoderma*, vol. 310, pp. 29–43, 2018.

[276] Xu *et al.*, "Assessment of Important Soil Properties Related to Chinese Soil Taxonomy Based on Vis–NIR Reflectance Spectroscopy," *Computers and Electronics in Agriculture*, vol. 144, pp. 1–8, 2018.

[277] Ng *et al.*, "In Search of an Optimum Sampling Algorithm for Prediction of Soil Properties From Infrared Spectra," *PeerJ*, vol. 6, p. e5722, 2018.

[278] Qi *et al.*, "Evaluating Calibration Methods for Predicting Soil Available Nutrients Using Hyperspectral VNIR Data," *Soil and Tillage Research*, vol. 175, pp. 267–275, 2018.

[279] B. Miloš and A. Bensa, "Prediction of Soil Organic Carbon Using Vis-NIR Spectroscopy: Application to Red Mediterranean Soils From Croatia," *Eurasian Journal of Soil Science*, vol. 6, no. 4, pp. 365–373, 2017.

[280] Jia *et al.*, "Hyperspectral Imaging Analysis for the Classification of Soil Types and the Determination of Soil Total Nitrogen," *Sensors*, vol. 17, no. 10, p. 2252, 2017.

[281] Zhang *et al.*, "Soil Nitrogen Content Forecasting Based on Real-Time Nir Spectroscopy," *Computers and Electronics in Agriculture*, vol. 124, pp. 29–36, 2016.

[282] Morellos *et al.*, "Machine Learning Based Prediction of Soil Total Nitrogen, Organic Carbon and Moisture Content by Using Vis-NIR Spectroscopy," *Biosystems Engineering*, vol. 152, pp. 104–116, 2016.

[283] Srivastava *et al.*, "Development of Hyperspectral Model for Rapid Monitoring of Soil Organic Carbon Under Precision Farming in the Indo-Gangetic Plains of Punjab, India," *Journal of the Indian Society of Remote Sensing*, vol. 43, pp. 751–759, 2015.

[284] M. Vohland *et al.*, "Determination of Soil Properties With Visible to Near- and Mid-infrared Spectroscopy: Effects of Spectral Variable Selection," *Geoderma*, vol. 223-225, pp. 88–96, 2014.

[285] Nocita *et al.*, "Prediction of Soil Organic Carbon Content by Diffuse Reflectance Spectroscopy Using a Local Partial Least Square Regression Approach," *Soil Biology and Biochemistry*, vol. 68, pp. 337–347, 2014.

[286] B. Sheeba *et al.*, "Machine Learning Algorithm for Soil Analysis and Classification of Micronutrients in IoT-Enabled Automated Farms," *Journal of Nanomaterials*, vol. 2022, 2022.

[287] Qi *et al.*, "A Novel Model for Soil Organic Matter and Total Nitrogen Detection Based on Visible/Shortwave Near-Infrared Spectroscopy," *Land*, vol. 14, no. 2, p. 329, 2025.

[288] Li *et al.*, "Multi-Scale Spatial and Spectral Feature Fusion for Soil Carbon Content Prediction Based on Hyperspectral Images," *Ecological Indicators*, vol. 160, p. 111843, 2024.

[289] Samarinas *et al.*, "Soil Data Cube and Artificial Intelligence Techniques for Generating National-Scale Topsoil Thematic Maps: A Case Study in Lithuanian Croplands," *Remote Sensing*, vol. 15, no. 22, p. 5304, 2023.

[290] Gao *et al.*, "Multivariate Prediction of Soil Aggregate-associated Organic Carbon by Simulating Satellite Sensor Bands," *Computers and Electronics in Agriculture*, vol. 209, p. 107859, 2023.

[291] Orgiazzi *et al.*, "Lucas Soil, the Largest Expandable Soil Dataset for Europe: A Review," *European Journal of Soil Science*, vol. 69, no. 1, pp. 140–153, 2018.

[292] Liu *et al.*, "Digital Mapping of Soil Organic Carbon Density Using Newly Developed Bare Soil Spectral Indices and Deep Neural Network," *Catena*, vol. 219, p. 106603, 2022.

[293] Wang *et al.*, "Hyperspectral Estimation of Soil Copper Concentration Based on Improved TabNet Model in the Eastern Junggar Coalfield," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[294] Devidhanshrii *et al.*, "Multiobjective Convolution Neural Network Towards Soil Nutrients Classification for Crop Recommendation on Based on Spectral and Spatial Properties Using Landsat Hyperspectral Images," *Journal of Pharmaceutical Negative Results*, pp. 2021–2031, 2022.

[295] Pham *et al.*, "Soil Profile Analysis Using Interactive Visualizations, Machine Learning, and Deep Learning," *Computers and Electronics in Agriculture*, vol. 191, p. 106539, 2021.

[296] Tsimpouris *et al.*, "Using Autoencoders to Compress Soil Vnir–Swir Spectra for More Robust Prediction of Soil Properties," *Geoderma*, vol. 393, p. 114967, 2021.

[297] Wang *et al.*, "A Deep Learning-based Method for Screening Soil Total Nitrogen Characteristic Wavelengths," *Computers and Electronics in Agriculture*, vol. 187, p. 106228, 2021.

[298] Patel *et al.*, "Deep-learning-based Approach for Estimation of Fractional Abundance of Nitrogen in Soil From Hyperspectral Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6495–6511, 2020.

[299] Reda *et al.*, "Predicting Soil Phosphorus and Studying the Effect of Texture on the Prediction Accuracy Using Machine Learning Combined With Near-Infrared Spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 242, p. 118736, 2020.

[300] Zhang *et al.*, "Temporal Paradox in Soil Potassium Estimations Using Spaceborne Multispectral Imagery," *Catena*, vol. 194, p. 104771, 2020.

[301] Tsakiridis *et al.*, "Simultaneous Prediction of Soil Properties From Vnir-Swir Spectra Using a Localized Multi-Channel 1-D Convolutional Neural Network," *Geoderma*, vol. 367, p. 114208, 2020.

[302] Yang *et al.*, "Combination of Convolutional Neural Networks and Recurrent Neural Networks for Predicting Soil Properties Using Vis–NIR Spectroscopy," *Geoderma*, vol. 380, p. 114616, 2020.

[303] Padarian *et al.*, "Using Deep Learning to Predict Soil Properties From Regional Spectral Data," *Geoderma Regional*, vol. 16, p. e00198, 2019.

[304] Fernandes *et al.*, "Estimation of Soil Organic Matter Content by Modeling With Artificial Neural Networks," *Geoderma*, vol. 350, pp. 46–51, 2019.

[305] Ng *et al.*, "The Influence of Training Sample Size on the Accuracy of Deep Learning Models for the Prediction of Soil Properties With Near-Infrared Spectroscopy Data," *Soil*, vol. 6, no. 2, pp. 565–578, 2020.

[306] Ng *et al.*, "Convolutional Neural Network for Simultaneous Prediction of Several Soil Properties Using Visible/Near-Infrared, Mid-Infrared, and Their Combined Spectra," *Geoderma*, vol. 352, pp. 251–267, 2019.

[307] Tsakiridis *et al.*, "A Genetic Algorithm-Based Stacking Algorithm for Predicting Soil Organic Matter From Vis-NIR Spectral Data," *European Journal of Soil Science*, vol. 70, no. 3, pp. 578–590, 2019.

[308] Tsakiridis *et al.*, "An Evolutionary Fuzzy Rule-Based System Applied to the Prediction of Soil Organic Carbon From Soil Spectral Libraries," *Applied Soft Computing*, vol. 81, p. 105504, 2019.

[309] Nikolaos *et al.*, "Using Interpretable Fuzzy Rule-Based Models for the Estimation of Soil Organic Carbon From Vnir/Swir Spectra and Soil Texture," *Chemometrics and Intelligent Laboratory Systems*, vol. 189, pp. 39–55, 2019.

[310] Singh *et al.*, "Estimation of Soil Properties From the EU Spectral Library Using Long Short-Term Memory Networks," *Geoderma Regional*, vol. 18, p. e00233, 2019.

[311] Wang *et al.*, "New Methods for Improving the Remote Sensing Estimation of Soil Organic Matter Content (SOMC) in the Ebinur Lake Wetland National Nature Reserve (ELWNNR) in Northwest China," *Remote Sensing of Environment*, vol. 218, pp. 104–118, 2018.

[312] Sila *et al.*, "Evaluating the Utility of Mid-infrared Spectral Subspaces for Predicting Soil Properties," *Chemometrics and Intelligent Laboratory Systems*, vol. 153, pp. 92–105, 2016.

[313] Zayani *et al.*, "Using Machine-Learning Algorithms to Predict Soil Organic Carbon Content from Combined Remote Sensing Imagery and Laboratory Vis-NIR Spectral Datasets," *Remote Sensing*, vol. 15, no. 17, p. 4264, 2023.

[314] Sun *et al.*, "Soil Knowledge-Guided Multi-Task Transformer Model for Predicting Soil Properties and Crop Traits With Early Warning Alerts," *Artificial Intelligence in Agriculture*, 2026.

[315] Li *et al.*, "Soil Carbon, Nitrogen and Phosphorus Contents Prediction Based on Hyperspectral Images Combined With Attention Mechanism Multi-Scale Dual-Branch Network," *Microchemical Journal*, p. 116380, 2025.

[316] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data With Neural Networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[317] Xu *et al.*, "Deep Learning Application for Predicting Soil Organic Matter Content by Vis-NIR Spectroscopy," *Computational Intelligence and Neuroscience*, vol. 2019, pp. 1–11, 2019.

[318] "Bhuvan HySI," https://www.geo.university/pages/spectral-indices-in-remote-sensing-and-how-to-interpret-them, Last accessed on October 8, 2023.

[319] A. A. Gitelson, "Wide Dynamic Range Vegetation Index for Remote Quantification of Biophysical Characteristics of Vegetation," *Journal of Plant Physiology*, vol. 161, no. 2, pp. 165–173, 2004.

[320] Nguyen *et al.*, "A Modified Bare Soil Index to Identify Bare Land Features During Agricultural Fallow-period in Southeast Asia Using Landsat 8," *Land*, vol. 10, no. 3, p. 231, 2021.

[321] Schut *et al.*, "Improved Wheat Yield and Production Forecasting With a Moisture Stress Index, Avhrr and Modis Data," *Crop and Pasture Science*, vol. 60, no. 1, pp. 60–70, 2009.

[322] Morabito *et al.*, "Environmental Temperature and Thermal Indices: What is the Most Effective Predictor of Heat-related Mortality in Different Geographical Contexts," *The Scientific World Journal*, vol. 2014, 2014.

[323] Wang *et al.*, "Remote Sensing Prediction Model of Cultivated Land Soil Organic Matter Considering the Best Time Window," *Sustainability*, vol. 15, no. 1, p. 469, 2022.

[324] Yu *et al.*, "Estimating and Mapping of Soil Organic Matter Content in a Typical River Basin of the Qinghai-Tibet Plateau," *Geocarto International*, vol. 37, no. 14, pp. 4088–4107, 2022.

[325] Yu *et al.*, "Improving Estimation of Soil Organic Matter Content by Combining Landsat 8 Oli Images and Environmental Data: A Case Study in the River Valley of the Southern Qinghai-Tibet Plateau," *Computers and Electronics in Agriculture*, vol. 185, p. 106144, 2021.

[326] John *et al.*, "Using Machine Learning Algorithms to Estimate Soil Organic Carbon Variability With Environmental Variables and Soil Nutrient Indicators in an Alluvial Soil," *Land*, vol. 9, no. 12, p. 487, 2020.

[327] Kumar *et al.*, "Geospatial Mapping of Soil Organic Carbon Using Regression Kriging and Remote Sensing," *Journal of the Indian Society of Remote Sensing*, vol. 46, pp. 705–716, 2018.

[328] Xu *et al.*, "Estimating Soil Total Nitrogen in Smallholder Farm Settings Using Remote Sensing Spectral Indices and Regression Kriging," *Catena*, vol. 163, pp. 111–122, 2018.

[329] Li *et al.*, "Machine Learning for Modeling Soil Organic Carbon as Affected by Land Cover Change in the Nebraska Sandhills, USA," *Environmental Modeling & Assessment*, vol. 29, no. 3, pp. 535–547, 2024.

[330] Liu *et al.*, "Prediction of the Soil Organic Carbon in the LUCAS Soil Database Based on Spectral Clustering." *Soil & Water Research*, vol. 18, no. 1, 2023.

[331] Gouda *et al.*, "Artificial Intelligence-Based Prediction of Key Textural Properties from LUCAS and ICRAF Spectral Libraries," *Agronomy*, vol. 11, no. 8, p. 1550, 2021.

[332] Pflugmacher *et al.*, "Mapping Pan-European Land Cover using Landsat Spectral-Temporal Metrics and the European LUCAS Survey," *Remote Sensing of Environment*, vol. 221, pp. 583–595, 2019.

[333] TOTH *et al.*, *LUCAS Topoil Survey-Methodology, Data and Results*, 2013, Last accessed on September 20, 2025. [Online]. Available: https://publications.jrc.ec.europa.eu/repository/handle/JRC83529

[334] "What is arcgis?" https://gisgeography.com/what-is-arcgis/, Last accessed on September 24, 2023.

[335] N. Flood, "Continuity of Reflectance Data Between Landsat-7 ETM+ and Landsat-8 OLI, for Both Top-of-atmosphere and Surface Reflectance: A Study in the Australian Landscape," *Remote Sensing*, vol. 6, no. 9, pp. 7952–7970, 2014.

[336] "Landsat 8 Bands," https://landsat.gsfc.nasa.gov/satellites/landsat-8/landsat-8-bands/, 2023, Last accessed on September 25, 2025.

[337] Izquierdo *et al.*, "Radiomics-Based Classification of Left Ventricular Non-Compaction, Hypertrophic Cardiomyopathy, and Dilated Cardiomyopathy in Cardiovascular Magnetic Resonance," *Frontiers in Cardiovascular Medicine*, vol. 8, p. 764312, 2021.

[338] Xing *et al.*, "The Correlation Between Soil Nutrient and Potato Quality in Loess Plateau of China based on PLSR," *Sustainability*, vol. 12, no. 4, p. 1588, 2020.

[339] Pasolli *et al.*, "Estimating Soil Moisture with the Support Vector Regression Technique," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 6, pp. 1080–1084, 2011.

[340] Kashi *et al.*, "Estimation of Soil Infiltration and Cation Exchange Capacity Based on Multiple Regression, ANN (RBF, MLP), and ANFIS Models," *Communications in Soil Science and Plant Analysis*, vol. 45, no. 9, pp. 1195–1213, 2014.

[341] Barkataki *et al.*, "Estimation of Soil Moisture from GPR Data using Artificial Neural Networks," in *IEEE Proc. International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES)*, 2021, pp. 1–5.

[342] J. Li, "Assessing the Accuracy of Predictive Models for Numerical Data: Not R nor R2, Why Not? Then What?" *PloS One*, vol. 12, no. 8, p. e0183250, 2017.

[343] Miaou *et al.*, "Pitfalls of Using R2 To Evaluate Goodness of Fit of Accident Prediction Models," *Transportation Research Record*, vol. 1542, no. 1, pp. 6–13, 1996.

[344] Tyagi *et al.*, "Nondestructive Identification of Wheat Species using Deep Convolutional Networks with Oversampling Strategies on Near-Infrared Hyperspectral Imagery," *Journal of Nondestructive Evaluation*, vol. 44, no. 1, pp. 1–21, 2025.

[345] Pang *et al.*, "Deep-Learning-Based Identification of Broad-Absorption Line Quasars," *Applied Sciences*, vol. 15, no. 3, p. 1024, 2025.

[346] Shi *et al.*, "State Space Models Meet Transformers for Hyperspectral Image Classification," *Signal Processing*, vol. 226, p. 109669, 2025.

[347] L. Kühnlein and S. Keller, "Evaluation Of Transformers And Convolutional Neural Networks For High-Dimensional Hyperspectral Soil Texture Classification," in *IEEE Proc. 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2022, pp. 1–5.

[348] Jin *et al.*, "An Innovative Approach for Integrating Two-dimensional Conversion of Vis-NIR Spectra With the Swin Transformer Model to Leverage Deep Learning for Predicting Soil Properties," *Geoderma*, vol. 436, p. 116555, 2023.

[349] Cao *et al.*, "A Novel Transformer-CNN Approach for Predicting Soil Properties From LUCAS Vis-NIR Spectral Data," *Agronomy*, vol. 14, no. 9, p. 1998, 2024.

[350] Zhai *et al.*, "PS-MTL-LUCAS: A Partially Shared Multi-Task Learning Model for Simultaneously Predicting Multiple Soil Properties," *Ecological Informatics*, vol. 82, p. 102784, 2024.

[351] Barnes *et al.*, "Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra," *Applied Spectroscopy*, vol. 43, no. 5, pp. 772–777, 1989.

[352] Sun *et al.*, "Novel Gumbel-Softmax Trick Enabled Concrete Autoencoder With Entropy Constraints for Unsupervised Hyperspectral Band Selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[353] Xu *et al.*, "Dropout Concrete Autoencoder for Band Selection on Hyperspectral Image Scenes," *IEEE Geoscience and Remote Sensing Letters*, vol. 22, pp. 1–5, 2025.

[354] Akande *et al.*, "Improving Soil Property Mapping Using Support Vector Machines, Neural Networks, Gradient Boosted Trees and Random Forests Over Soils in Sub-Saharan Africa," *Neural Networks, Gradient Boosted Trees and Random Forests Over Soils in Sub-Saharan Africa*, 2022.

[355] Dlamini *et al.*, "Overgrazing Decreases Soil Organic Carbon Stocks the Most Under Dry Climates and Low Soil pH: A Meta-analysis Shows," *Agriculture, Ecosystems & Environment*, vol. 221, pp. 258–269, 2016.

[356] Chu *et al.*, "Data Cleaning: Overview and Emerging Challenges," in *Proc. International Conference on Management of Data*, 2016, pp. 2201–2206.

[357] Reddy *et al.*, "Maintenance Enhancement of Smart Manufacturing Units In Industry 4.0," in *IEEE Proc. International Conference on Inventive Computation Technologies (ICICT)*, 2023, pp. 1488–1492.

[358] C. Ferreira, "Gene Expression Programming in Problem Solving," in *Soft Computing and Industry: Recent Applications*. Springer, 2002, pp. 635–653.

[359] A. Maćkiewicz and W. Ratajczak, "Principal Components Analysis (PCA)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[360] V. Fonti and E. Belitser, "Feature Selection Using Lasso," *VU Amsterdam Research Paper in Business Analytics*, vol. 30, pp. 1–25, 2017.

[361] Soares *et al.*, "The Successive Projections Algorithm," *TrAC Trends in Analytical Chemistry*, vol. 42, pp. 84–98, 2013.

[362] Darst *et al.*, "Using Recursive Feature Elimination in Random Forest to Account for Correlated Variables in High Dimensional Data," *BMC Genetics*, vol. 19, pp. 1–6, 2018.

[363] Zhang *et al.*, "Comparison of New Hyperspectral Index and Machine Learning Models for Prediction of Winter Wheat Leaf Water Content," *Plant Methods*, vol. 17, pp. 1–14, 2021.

Environ Monit Assess (2024) 196:699
https://doi.org/10.1007/s10661-024-12817-6

REVIEW

A critical systematic review on spectral-based soil nutrient prediction using machine learning

Shagun Jain · Divyashikha Sethia · Kailash Chandra Tiwari

Received: 14 April 2024 / Accepted: 11 June 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

Abstract The United Nations (UN) emphasizes the pivotal role of sustainable agriculture in addressing persistent starvation and working towards zero hunger by 2030 through global development. Intensive agricultural practices have adversely impacted soil quality, necessitating soil nutrient analysis for enhancing farm productivity and environmental sustainability. Researchers increasingly turn to Artificial Intelligence (AI) techniques to improve crop yield estimation and optimize soil nutrition management. This study reviews 155 papers published from 2014 to 2024, assessing the use of machine learning (ML) and deep learning (DL) in predicting soil nutrients. It highlights the potential of hyperspectral and multispectral sensors, which enable precise nutrient identification through spectral analysis across multiple bands. The study underscores the importance of feature selection techniques to improve model performance by eliminating redundant spectral bands with weak correlations to targeted nutrients. Additionally, the use of spectral indices, derived from mathematical ratios of spectral bands based on absorption spectra, is examined for its effectiveness in accurately predicting soil nutrient levels. By evaluating various performance measures and datasets related to soil nutrient prediction, this paper offers comprehensive insights into the applicability of AI techniques in optimizing soil nutrition management. The insights gained from this review can inform future research and policy decisions to achieve global development goals and promote environmental sustainability.

Keywords Sustainable agriculture · Soil nutrients · Hyperspectral · Machine learning · Deep learning

S. Jain (✉) · D. Sethia
Department of Software Engineering, Delhi Technological University, Delhi, India
e-mail: shagunjain191172@gmail.com

D. Sethia
e-mail: divyashikha@dtu.ac.in

K. C. Tiwari
Multidisciplinary Centre of Geoinformatics, Delhi Technological University, Delhi, India
e-mail: kcchtphd@gmail.com

Introduction

The United Nations's (UN's) second Sustainable Development Goal (SDG) is zero hunger, aiming to promote sustainable agriculture practices. The report highlights the necessity of transforming food systems and addressing the pandemic's impact on global nutrition to tackle increasing challenges in food security (United Nations, The Sustainable Development Goals Report, 2023). However, despite efforts to ensure sufficient food production, crops can suffer from various issues, such as damage and diseases. Surprisingly, these effects are visible in crops but not in the soil itself, which does not exist independently but in a mixed form with vegetation and other substances (Chandraprabha et al., 2021). Global evidence suggests that variations in soil nutrient levels immediately affect food production,

Published online: 04 July 2024

Environ Monit Assess (2024) 196:1255
https://doi.org/10.1007/s10661-024-13406-3

RESEARCH

Developing novel spectral indices for precise estimation of soil pH and organic carbon with hyperspectral data and machine learning

Shagun Jain · Divyashikha Sethia · Kailash Chandra Tiwari

Received: 22 July 2024 / Accepted: 12 November 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

Abstract Accurate soil pH and soil organic carbon (SOC) estimations are vital for sustainable agriculture, as pH affects nutrient availability, and SOC is crucial for soil health and fertility. Hyperspectral imaging provides a faster, non-destructive, and economical alternative to standard soil testing. The study utilizes imaging spectroscopic data from the Africa Soil Information Service (AfSIS) and Land Use and Coverage Area Frame Survey (LUCAS-2009) hyperspectral datasets, capturing spatially distributed spectral information. Machine learning (ML) approaches using high-dimensional spectral bands can be computationally expensive, while those using spectral indices are typically limited to multispectral data. This study addresses these challenges by comparing soil pH and SOC prediction using ML models, with both existing spectral indices and individual hyperspectral bands as input features. Results demonstrate that hyperspectral bands outperform existing indices in predictive accuracy, with R^2 values ranging from 0.8 to 0.94 for both soil pH and SOC. To further enhance prediction performance, this study proposes novel spectral indices-soil pH index (SPI) and soil organic carbon index (SOCI)-specifically designed for hyperspectral data using principal component analysis (PCA) and artificial neural networks (ANN). The proposed SPI and SOCI indices address multicollinearity issues and high dimensionality in raw spectral bands, significantly improving predictive accuracy. The SPI and SOCI indices achieve R^2 values of 0.86 for AfSIS soil pH, 0.945 for LUCAS-2009 soil pH, 0.952 for AfSIS SOC, and 0.963 for LUCAS-2009 SOC. These results show that the proposed spectral indices provide a practical solution for precision agriculture, enhancing soil pH and SOC estimations.

Keywords Machine learning · Spectral indices · Soil pH · Soil organic carbon · Hyperspectral · Sustainable agriculture

S. Jain (✉) · D. Sethia
Department of Software Engineering, Delhi Technological University, Delhi, India
e-mail: shagunjain191172@gmail.com

D. Sethia
e-mail: divyashikha@dtu.ac.in

K. C. Tiwari
Multidisciplinary Centre of Geoinformatics, Delhi Technological University, Delhi, India
e-mail: kcchtphd@gmail.com

Introduction

Maintaining healthy soil is paramount for achieving the sustainable development goals (SDGs) set by the United Nations (Awais et al., 2023). These goals encompass crucial targets like ensuring food security (SDG 2: Zero Hunger), mitigating climate change (SDG 13: Climate Action), and promoting healthy ecosystems (SDG 15: Life on Land) (Veenstra et al., 2024). Soil pH and organic carbon content are pivotal indicators of soil health, influencing plant growth, nutrient availability, and overall ecosystem function-

Published online: 26 November 2024

CAT-MTLNet: A Novel Deep Multitarget Framework for Hyperspectral Soil Properties Estimation Using Concrete Autoencoder and Transformer

Shagun Jain, Divyashikha Sethia, and Kailash Chandra Tiwari

***Abstract*—Due to their advanced feature extraction capabilities, deep learning (DL) models have shown significant success in estimating soil properties from hyperspectral data. However, challenges persist in accurately estimating multiple soil properties simultaneously while handling high-dimensional spectral inputs. Hence, this letter proposes a novel multitarget learning framework—concrete autoencoder and transformer-based multitarget learning network (CAT-MTLNet) to address these issues. It integrates a concrete autoencoder (CAE) for optimal spectral band selection with a transformer-multilayer perceptron (MLP) model for robust multisoil properties estimation. The framework jointly estimates soil organic carbon (SOC), phosphorus (P), pH, and calcium (Ca). Experimental evaluations on the Africa soil information service (AfSIS) and LUCAS-2009 datasets demonstrate the effectiveness of CAT-MTLNet. For AfSIS, R^2 scores range from 0.453 to 0.953 with root mean square error (RMSE) values between 0.289 and 0.390. For LUCAS-2009, R^2 scores range from 0.209 to 0.931, and RMSE varies from 26.094 to 39.006. CAT-MTLNet successfully balances interpretability, dimensionality reduction, and predictive performance, offering a lightweight and scalable solution for estimating hyperspectral soil properties.**

***Index Terms*—Attention mechanism, hyperspectral imaging, multioutput regression, soil properties estimation, soil spectroscopy, spectral band selection.**

I. Introduction

SOIL plays a fundamental role in agricultural productivity, environmental health, and food security [1]. The balanced presence of key soil properties, such as soil organic carbon (SOC), phosphorus (P), pH, and calcium (Ca), is critical for sustainable crop production [2]. Hyperspectral spectroscopy, particularly in the visible and near-infrared (Vis–NIR) range, has emerged as a powerful, nondestructive tool for soil properties assessment due to its ability to capture hundreds of narrow, contiguous spectral bands that reveal subtle chemical characteristics [3].

The Africa soil information service (AfSIS) [4] and LUCAS-2009 [5] datasets are two widely adopted benchmarks that provide hyperspectral reflectance data paired with reference measurements of essential soil properties. These datasets differ in geographical coverage, spectral resolution, and sample diversity, making them suitable benchmarks for developing generalized and scalable predictive models.

Recent advances in deep learning (DL) have enabled more effective modeling of hyperspectral soil data. Convolutional neural networks (CNNs) are widely used for their ability to capture local spectral patterns [6], while transformer architectures provide global contextual modeling through self-attention mechanisms [7]. Various DL models have proposed architectures that use the combinations of CNN, transformer, and attention layer, such as multitarget CNNs with attention modules [8], CNN–transformer hybrids [9], image-based Swin Transformers [10], and partially shared multitarget learning frameworks with post-hoc interpretation [11]. As summarized in Table I, these state-of-the-art (SOTA) approaches rely heavily on full-spectrum inputs and deep architectures, leading to high computational demands and limited interpretability. None of the prior models integrates feature selection into the training process, and most use CNN-based regressors that increase complexity and resource requirements. This underscores the importance of incorporating feature selection to identify the most informative spectral bands.

Hence, this letter aims to develop a novel, unified, and scalable framework called concrete autoencoder and transformer-based multitarget learning network (CAT-MTLNet) that addresses the computational, interpretability, and scalability limitations of previous models [8], [9], [10], [11]. It integrates a concrete autoencoder (CAE) to perform target-aware, differentiable spectral band selection, thereby reducing input redundancy and enhancing interpretability. A transformer encoder models global spectral relationships, and a lightweight multilayer perceptron (MLP) performs multioutput regression, reducing parameter overhead compared to CNNs, making the model efficient.

The key contributions of this letter include the following.

1) Proposal of a novel, lightweight, interpretable, and scalable framework, CAT-MTLNet—a CAE and transformer-MLP based multitarget learning network for simultaneous SOC, P, pH, and Ca estimation using hyperspectral data.
2) Evaluation of the CAT-MTLNet on two benchmark datasets—AfSIS and LUCAS-2009—along with a comparison of CAE against attention-guided feature pruning (AFGP) and reinforcement learning (RL), demonstrates its competitive accuracy, lower complexity, and strong generalizability.

Received 8 September 2025; revised 17 October 2025; accepted 17 November 2025. Date of publication 20 November 2025; date of current version 4 December 2025. *(Corresponding author: Shagun Jain.)*

Shagun Jain and Divyashikha Sethia are with the Department of Software Engineering, Delhi Technological University, New Delhi 110042, India (e-mail: shagunjain191172@gmail.com; divyashikha@dtu.ac.in).

Kailash Chandra Tiwari is with the Multidisciplinary Centre of Geoinformatics, Delhi Technological University, New Delhi 110042, India (e-mail: kcchtphd@gmail.com).

Digital Object Identifier 10.1109/LGRS.2025.3635157

1558-0571 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Vivekananda Institute of Professional Studies. Downloaded on December 08,2025 at 11:05:31 UTC from IEEE Xplore. Restrictions apply.

Computers and Electronics in Agriculture 243 (2026) 111381

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

TLM-stack: A deep learning-based novel framework for soil nutrients estimation using hyperspectral data

Shagun Jain [a],*, Divyashikha Sethia [a], Kailash Chandra Tiwari [b]

[a] Department of Software Engineering, Delhi Technological University, Delhi, India
[b] Multidisciplinary Centre of Geoinformatics, Delhi Technological University, Delhi, India

ARTICLE INFO

Keywords:
Soil nutrients
Hyperspectral data
Feature selection
Deep learning
Stacking ensemble learning

ABSTRACT

Soil analysis provides essential information necessary to improve soil quality. A farmer may quickly modify fertilization to meet crop and soil requirements by monitoring the precise amount of nutrients in the soil. Developing a rapid and effective estimation model using hyperspectral remote sensing data is essential for the widespread implementation of soil nutrient inversion mapping across extensive regions. This paper evaluates and compares the estimated potentials of multiple Deep Learning (DL) algorithms and proposes a novel DL-based stacking ensemble framework, TLM-Stack (TabNet-LSTM-MLP), for the estimation of numerous soil nutrients, including Soil Organic Carbon (SOC), Phosphorous (P), soil pH, Nitrogen (N) and Calcium (Ca) using hyperspectral data for improved performance. The study also evaluates various feature selection methods to determine the optimal spectral feature subset. The proposed TLM-Stack framework is validated on heavy datasets, such as the Land Use and Coverage Area Frame Survey (LUCAS-2009) and the Africa Soil Information Service (AfSIS), consistently demonstrating accuracy and efficiency in real-world applications. Results indicate that combining Principal Component Analysis (PCA) with the proposed TLM-Stack model achieves optimal estimation for SOC, P, pH, N, and Ca, with testing R^2 values ranging from 0.46 to 0.97, respectively. The DL-based estimation model, combined with the strategy of base and meta-learners, exhibits better predictive effects and stability than the single algorithm, thereby reducing the risk of overfitting. The feature selection method makes the proposed TLM-Stack framework lightweight, space-efficient, and ideal for estimating soil nutrients.

1. Introduction

Soil is widely recognized as an essential component of the environment in which humans live and is the basic structural component of the biosphere (Diaz-Gonzalez et al., 2022; Andrés-Domenech et al., 2023). It is essential for converting solar energy into biological energy. Soil is also a natural resource that humans, animals, and plants need, especially in production (Das et al., 2015). The nutrients in soil play a significant role in its effectiveness because they are necessary for plants to grow, develop, and reproduce (Jamei et al., 2024; Yu et al., 2021). Sustainable agricultural operations and environmental well-being require an understanding and regulation of soil nutrients (Srivastava et al., 2021; Zhou et al., 2021). Measuring the nutrient content of soil requires tests to assess its health (Pham et al., 2021). However, conventional methods of analyzing soil nutrients are time-consuming, labor-intensive, and often necessitate fieldwork (Gulhane et al., 2023).

Covering extensive geographic areas is essential for effective soil analysis. Remote sensing technologies enable comprehensive soil assessments over large regions, providing accurate forecasts of soil nutrient levels through advanced data analytics (Roy and George K, 2020). Hyperspectral sensors capture detailed spectral information across continuous bands, revealing various surface properties (Keller et al., 2018; Malik et al., 2023). However, the high number of spectral channels often leads to redundant data and collinearity, which can degrade the overall quality of the input data (Wang et al., 2018).

The Land Use and Coverage Area Survey (LUCAS-2009) (Fernández-Ugalde et al., 2020) and Africa Soil Information Service (AfSIS) (Leenaars, 2013) datasets serve as critical resources in this study. Both datasets include laboratory-analyzed spectral data and detailed mineral and organic soil compositions, making them ideal for soil nutrient estimation tasks (Qiao et al., 2022; Gruszczyński et al., 2022). LUCAS-2009 covers extensive soil samples across Europe, while AfSIS focuses on African soils, reflecting diverse geographical and environmental

* Corresponding author.
E-mail addresses: shagunjain191172@gmail.com (S. Jain), divyashikha@dtu.ac.in (D. Sethia), kcchtphd@gmail.com (K.C. Tiwari).

https://doi.org/10.1016/j.compag.2025.111381
Received 5 November 2024; Received in revised form 16 December 2025; Accepted 23 December 2025
Available online 3 January 2026
0168-1699/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

A Review on Applications of Artificial Intelligence for Identifying Soil Nutrients

Shagun Jain(✉) and Divyashikha Sethia

Department of Software Engineering, Delhi Technological University, Delhi, India
shagunjain191172@gmail.com, divyashikha@dtu.ac.in

Abstract. Soil nutrients play a crucial role in the growth and productivity of crops, which are essential for meeting the increasing global food demand. However, the traditional methods of soil nutrient analysis are time-consuming, labour-intensive, and often not cost-effective. Recently, there has been a growing interest in applying Artificial Intelligence (AI) techniques to soil nutrient analysis. AI can help optimize soil nutrient management and improve crop yields. Farmers can identify potential nutrient deficits in soil quality using artificial intelligence technology, particularly electronic applications for Machine Learning (ML) and Deep Learning (DL). This paper aims to summarise and assess state of the art in using Artificial Intelligence for soil nutrient analysis. The study explores the different ML and DL techniques used for soil nutrient analysis and the results obtained from their application. The review additionally points out areas of improvement for the current research and suggests possibilities for future investigations. It further explains different datasets based on satellite images, smartphone images, and chemical data. Furthermore, the report includes several publicly available soil datasets.

Keywords: Soil nutrients · Artificial Intelligence · Machine Learning · Deep Learning

1 Introduction

Soil is a vital resource for the Earth but still needs to be appreciated. The health of the soil has significantly impacted the state of the food chain, the quality of air and water, and many other ecosystems on Earth. Earth's population is overgrowing, and soil nutrient analysis is essential to feed such a vast population in precision farming.

Soil nutrients heavily influence the yield of crops [3]. The primary source of nutrient absorption for plants is the soil. By understanding the nutrient status of the soil, farmers can take appropriate measures to adjust the levels of deficient nutrients and maintain optimal soil conditions for plant growth. It can help increase crop yield and improve the quality and consistency of the crops. Soil nutrient prediction improves agricultural productivity, environmental sustainability, and food security. Farmers, agronomists, and soil scientists

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
M. K. Saini et al. (Eds.): ICA 2023, CCIS 1866, pp. 71–86, 2023.
https://doi.org/10.1007/978-3-031-43605-5_6

2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI) | 979-8-3503-4279-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAEECI58247.2023.10370850

A Hybrid Approach for Soil Nutrient Estimation using Multispectral Data in Belgium and Luxembourg

Shagun Jain[a*], Divyashikha Sethia[b], and K.C. Tiwari[c]

[a*,b]Department of Software Engineering, Delhi Technological University, New Delhi, India
[c]Department of Civil Engineering, Delhi Technological University, New Delhi, India
[a*]shagunjain191172@gmail.com

***Abstract*—Planning for agricultural production and sustainable land management depend on accurately measuring soil nutrient levels. Leveraging machine learning to estimate soil nutrients simplifies farming and land management decision-making. This paper presents a novel hybrid model to forecast soil nutrient concentrations in the Belgium and Luxembourg region by combining Partial least squares regression (PLSR) with Gaussian process regression (GPR). An integrated approach is provided for predicting various soil nutrients, utilizing the Lucas topsoil dataset and multispectral Landsat 8 satellite images, extending the model's applicability to numerous agricultural and environmental applications. The hybrid model has undergone evaluation, producing RMSE and RPD values within acceptable ranges for various soil parameters, such as Soil Organic Carbon (SOC), Electrical Conductivity (EC), Nitrogen (N), Potassium (K), Phosphorus (P), pH in Water (H2O), pH in Calcium Chloride (CaCl2), and pH in Calcium Carbonate (CaCO3).**

***Index Terms*—Machine Learning, Multispectral, Soil nutrients, Artificial Intelligence, Lucas top soil dataset**

I. Introduction

Soil is the foundation for sustaining plant life and fostering agricultural output and is a crucial element of the terrestrial ecosystem. Its functionality mostly depends on the soil's nutrients, which plants need for average growth, development, and reproduction. Soil nutrient management and understanding are crucial for sustaining sustainable agricultural and environmental well-being [1]. It is essential to test the soil health for measuring the soil nutrients. However, traditional soil testing techniques frequently involve fieldwork, are time-consuming, and demand much labor.

Machine learning, a subset of artificial intelligence, has emerged as a powerful technique for estimating soil nutrients due to its capacity to extract information from extensive datasets and generate predictions. Jain et al. [2] have recently reviewed the emerging field of machine learning for assessing soil nutrients. Their study mentions various Artificial Intelligence (AI) models that predict soil nutrients. Making exact forecasts regarding the amounts of soil nutrients across enormous geographic areas is now possible because of recent advancements in data analytics and remote sensing [3].

The researchers utilize multispectral images from Landsat 8, obtained from the United States Geological Survey (USGS) Earth Explorer website. This valuable data source captures information across different wavelengths. Remote sensing platforms such as satellites provide multispectral data from multiple distinct spectral bands. These bands capture various surface features since they represent several electromagnetic spectrum regions. Multispectral data makes it possible to distinguish between different soil qualities in soil analysis based on their unique spectral fingerprints [4].

Castaldi et al. [5] focused their research on estimating a single nutrient—soil organic carbon (SOC)—within the Belgium and Luxembourg region. They used the Lucas topsoil dataset from 2012. In contrast, the present study aims to predict a broader array of soil nutrients using the Lucas topsoil dataset 2015. This expanded scope enhances the study's significance for agriculture and environmental applications.

This paper utilizes multispectral data to estimate soil nutrient concentrations for Belgium and Luxembourg. It aims to accelerate the assessment of soil nutrients by using machine learning and remote sensing, enabling agricultural professionals and amateurs to make well-informed decisions about crop choice and management. In particular, it focuses on predicting soil organic carbon, soil pH in water (H2O) and calcium chloride (CaCl2) solutions, calcium carbonate (CaCO3) content, electrical conductivity (EC), and the concentrations of Nitrogen (N), Phosphorous (P), and Potassium (K) in Lucas topsoil data from the year 2015 [6], which includes data from 28 different European nations, offer crucial ground level measurements that help in developing reliable prediction models [7].

The contributions of this paper are as follows:

- Developing a hybrid machine learning framework that uses acquired satellite images to estimate soil nutrients for Belgium and Luxembourg.
- Enabling the forecasting of various soil nutrients in the Belgium and Luxembourg regions.

The following sections go into more detail on several subjects. Section 2 provides context by reviewing previous studies. Section 3 details the research technique, including the data description, preprocessing steps, feature extraction, building of the machine learning model, and effectiveness

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on January 04,2024 at 07:57:12 UTC from IEEE Xplore. Restrictions apply.

Comparative Analysis of Machine Learning Based Soil pH Prediction Using Spectral Bands and Indices

Shagun Jain[1(✉)], Divyashikha Sethia[1], and Kailash Chandra Tiwari[2]

[1] Department of Software Engineering, Delhi Technological University, Delhi, India
shagunjain191172@gmail.com, divyashikha@dtu.ac.in
[2] Multidisciplinary Centre of Geoinformatics, Delhi Technological University, Delhi, India

Abstract. Soil pH is vital in agriculture and environmental science, affecting plant growth, nutrient availability, and soil health. Remote sensing and Machine Learning (ML) accurately predict and map soil pH, guiding effective soil management for optimized yields and environmental conservation. This paper compares ML-based soil pH prediction using multispectral bands and various spectral indices derived from remote sensing data. It uses indices such as vegetation indices, soil indices, temperature, and thermal indices individually and in combination to predict soil pH. Three primary ML models, namely Partial Least Square Regression (PLSR), Support Vector Regression (SVR), and Multilayer Perceptron (MLP), are utilized for predicting soil pH. The findings suggest that spectral indices yield marginally superior predictions compared to spectral bands alone. This finding highlights the significance of using spectral indices, especially vegetation indices, to estimate the soil pH. The incorporation of spectral indices improves the forecast precision because of the subtle data they record about the characteristics and composition of the soil. This research advances remote sensing in soil science by shedding light on the viability and effectiveness of using satellite data for soil pH prediction.

Keywords: Soil pH · Spectral bands · Spectral Indices · Machine Learning

1 Introduction

Soil pH significantly impacts soil health and the dynamics of entire ecosystems, making it a crucial factor in environmental science and agricultural activities [1]. It immediately affects the availability of nutrients, microbial activity, and soil structure [2]. For example, aluminium and manganese toxicity in acidic soils may prevent plants from growing. Conversely, micronutrient deficits may result from alkaline soils [3]. Therefore, knowing and controlling soil pH is essential for sustainable land usage and ideal crop production [4]. To maximize yields

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
M. K. Saini et al. (Eds.): ICA 2024, CCIS 2207, pp. 43–55, 2025.
https://doi.org/10.1007/978-3-031-74440-2_5

CNViS-Net: A Self-Attention Framework for Predicting Soil Nutrients from Hyperspectral Data

Shagun Jain[1], Divyashikha Sethia[1] and Kailash Chandra Tiwari[2]

[1]Department of Software Engineering, Delhi Technological University, Delhi, India
[2]Multidisciplinary Centre of Geoinformatics, Delhi Technological University, Delhi, India

E-mail : shagunjain191172@gmail.com, divyashikha@dtu.ac.in, kcchtphd@gmail.com

2025 International Conference on Frontier Technologies and Solutions (ICFTS) | 979-8-3315-0276-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICFTS62006.2025.11031492

Abstract-Automating soil parameter analysis simplifies fertilization, resulting in time and resource savings. This efficiency reduces agricultural production costs and fosters better resource utilization, contributing to sustainable farming practices and improving food security. This research presents CNViS-Net, a novel self-attention framework for soil nutrient prediction using hyperspectral data, combining Deep Learning (DL) with Machine Learning (ML) techniques. The focus is on predicting four soil nutrients: Phosphorus (P), Potassium (K), Magnesium (Mg), and soil pH, using the HYPERVIEW hyperspectral dataset. The proposed CNViS-Net framework performs feature extraction through a combination of Convolutional Neural Network (CNN) and Vision Transformer (ViT), capturing complex spectral patterns across 150 channels of hyperspectral data. It passes the extracted features to four ML models, which are K-Nearest Neighbors (KNN), Random Forest (RF), Partial Least Squares Regression (PLSR), and Support Vector Regression (SVR) for final prediction. The combination of SVR with CNN and ViT outperforms other prediction models, achieving significantly lower score metrics of 0.164, 0.074, 0.307, and 0.66 for P, K, Mg, and pH, respectively. These remarkably lower scores demonstrate CNViS-Net's superior performance over previous state-of-the-art methods in accurately predicting soil nutrient levels, advancing precision agriculture through efficient hyperspectral data analysis.

Index Terms—Convolutional neural network, Vision Transformer, Support vector regression, Hyperspectral, Soil Nutrients.

I. INTRODUCTION

Soil health is crucial for sustainable agriculture, with nutrient availability directly impacting crop success [1]. Accurate nutrient prediction helps optimize fertilizer use, boosting yields and promoting environmental sustainability [2]. Traditional lab-based soil analyses, though effective, are costly and impractical for large-scale, real-time use [3]. Hyperspectral imaging offers a powerful alternative, capturing detailed reflectance data across hundreds of spectral bands to precisely estimate soil nutrients, paving the way for efficient, large-scale soil health monitoring [4].

Hyperspectral data offers rich spectral detail but requires advanced processing [5]. Artificial Intelligence (AI) models effectively automate analysis, uncovering complex patterns in raw data without manual feature extraction [6]. These models reveal subtle spectral traits often missed by traditional methods, enabling efficient, precise, and scalable soil nutrient prediction.

Despite the potential of hyperspectral data in soil nutrient prediction, existing models often struggle with the complexity and high dimensionality of such datasets. Hyperspectral data contain numerous spectral bands, making feature extraction and model optimization challenging. Traditional Machine Learning (ML) techniques, while effective, can be computationally expensive and prone to overfitting when handling large datasets [7]. Additionally, manual feature extraction methods may fail to capture key spectral patterns, leading to reduced prediction accuracy [8]. Addressing these challenges requires advanced techniques that can efficiently process high-dimensional data while preserving essential spectral features for improved soil nutrient prediction.

Feature extraction helps in managing the high dimensionality of hyperspectral data, helping it to reduce computational load and improve model accuracy. Advanced methods like Vision Transformers (ViTs) have demonstrated strong potential in hyperspectral image analysis by capturing intricate spatial and spectral relationships within the data [9]. For instance, studies by Wang et al. [10] achieved high accuracy for urban ground object classification, mineral classification, and crop classification with ViT across multiple hyperspectral datasets, highlighting its effectiveness in addressing complex data contexts. Similarly, Kuhnlein et al. [11] successfully applied Convolutional Neural Networks (CNN) and ViT models for soil texture classification in the LUCAS dataset, further demonstrating ViT's capabilities in hyperspectral analysis. These studies underscore the adaptability of ViT in extracting meaningful features across varied hyperspectral datasets, showcasing their potential for robust classification across diverse application domains.

Recent advancements in ViTs, such as those in [10], [11], have shown great potential in hyperspectral imaging by

2024 International Conference on Augmented Reality, Intelligent Systems, and Industrial Automation (ARIIA) | 979-8-3315-1836-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ARIIA63345.2024.11051496

DSeP-xNet: A Feature Optimized Ensemble Framework for Hyperspectral-Based Soil Organic Carbon Prediction

Akshyat Shah
Department of Computer Science
Delhi Technological University
Delhi, India
akshyatshah_co22a1_58@dtu.ac.in

Shashi Sah
Department of Computer Science
Delhi Technological University
Delhi, India
shashisah_co22a1_54@dtu.ac.in

Shashvat Singhal
Department of Computer Science
Delhi Technological University
Delhi, India
shashvatsinghal_co22a7_21@dtu.ac.in

Shagun Jain
Department of Software Engineering Delhi Technological University Delhi, India
shagunjain191172@gmail.com

Divyashikha Sethi
Department of Software Engineering
Delhi Technological University
Delhi, India
divyashikha@dtu.ac.in

***Abstract*—Accurate measurement and management of soil nutrients are crucial for smart agriculture. Among various soil nutrients, Soil Organic Carbon (SOC) is one of the most essential measures that form the basis of this work. Unlike approaches that use individual Machine Learning (ML) and Deep Learning (DL) models with minimal emphasis on data cleaning and feature selection, this paper introduces a novel methodological framework called DSeP-xNet. The framework includes preprocessing techniques like noise removal with the Savitzky–Golay filter and feature selection using XGBoost, as minute changes in features using such techniques significantly affect the model accuracy. It also anchors around previously established models to configure a stacking-based ensemble model, DSeP-Net, combining the strengths of three base models: a Deep Neural Network (DNN), a Squeeze and Excitation Convolutional Neural Network (SECNN), and Partial Least Squares Regression (PLSR), with Linear Regression as the meta-model. The study uses data points from two publicly available hyperspectral imagery data: Africa Soil Information Service (AfSIS) and Land Use and Coverage Area Frame Survey (LUCAS-2009) for SOC prediction. The datasets leverage the individual pixels of satellite images over the narrow band of reflected light as features. The DSeP-xNet framework demonstrates a remarkable R^2 and RMSE score of 0.967 and 0.2202, respectively, on the AfSIS dataset and 0.9448 and 22.7193 on the LUCAS-2009 dataset, attesting to its high accuracy and robust nature in SOC prediction. These findings pave the way for more precise and efficient data-driven decision-making in agriculture and environmental conservation.**

***Index Terms*—Soil Organic Carbon, Feature Selection, Ensemble Technique, Hyperspectral, Artificial Intelligence**

I. Introduction

Soil nutrients are crucial in soil health, significantly influencing plant growth [1]. Assessing soil nutrients is essential for optimizing agricultural practices [2]. It enables the informed selection of appropriate supplements and fertilizers necessary to enhance farming yields, addressing the demands of a growing global population. Among these nutrients, Soil Organic Carbon (SOC) is a critical element directly linked to plant development [3]. SOC is the main component of Soil Organic Matter (SOM). It significantly affects microbial biomass and soil respiration, explaining about 58% of the observed variations.

The conventional approach to assessing soil nutrients involves manually collecting soil samples and sending them to a laboratory for testing [4]. The approaches use chemical laboratory procedures to provide researchers with the results and demand significant time and effort in return [5]. Furthermore, as soil quality can vary across expansive areas, more than a single soil sample is necessary for evaluating soil quality comprehensively and on a larger scale; these limitations underscore the necessity for more innovative techniques to gather soil nutrient data.

Besides traditional chemical methods employed by laboratories, modern techniques for soil nutrient estimation have emerged. These include optical and sensor-based methods, contributing to a more comprehensive approach to soil analysis [6]. Methods to measure soil with optical sensors include visible-infrared spectroscopy, reflectance spectroscopy, and colourimetric approaches. Spectroscopic methods analyze materials based on their energy absorption at specific wavelengths. Each soil nutrient's unique spectral signature facilitates its quantification and identification.

Using hyperspectral data for soil nutrient estimation has become increasingly popular. It involves capturing the reflected light in narrow bands across the electromagnetic spectrum [7]. This approach goes beyond the limitations of conventional manual and chemical methods, offering a rich dataset while enhancing the precision and sophistication of assessing soil nutrient levels using the extensive data range.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 02,2025 at 07:53:19 UTC from IEEE Xplore. Restrictions apply.

2026 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3315-4970-1/26/$31.00 ©2026 IEEE | DOI: 10.1109/IATMSI68868.2026.11465991

Multi-Strategy Fused Transformer-CNN with PSO-Driven Optimization for Soil Properties Estimation from Hyperspectral Data

Abhyudya Sangwan
Department of Software Engineering
Delhi Technological University
Delhi, India
work.abhyudya@gmail.com

Divyashikha Sethia
Department of Software Engineering
Delhi Technological University
Delhi, India
divyashikha@dtu.ac.in

Shagun Jain
Department of Software Engineering
Delhi Technological University
Delhi, India
shagunjain191172@gmail.com

***Abstract*—Accurate and scalable prediction of soil properties is critical for sustainable agriculture, enabling informed nutrient management and optimized land use. Key soil components such as Calcium (Ca), pH, Phosphorus (P), sand content, and Soil Organic Carbon (SOC) govern soil fertility, crop productivity, and environmental resilience. However, accurate prediction of these components with deep learning often remains constrained by reliance on a single architecture type or fixed fusion schemes. This paper presents a hyperspectral learning framework, Multi-Strategy Fused Transformer-CNN (MSF-TCNN), that integrates Particle Swarm Optimization (PSO) for spectral feature selection and hyperparameter tuning with a dual-branch deep neural architecture. The proposed model employs a Convolutional Neural Network (CNN) to extract localized absorption patterns from contiguous spectral regions, and a Transformer to capture long-range spectral dependencies. The paper employs four complementary fusion strategies (CNN-only, Transformer-only, direct concatenation of CNN and Transformer, and adaptive alpha-blend) to generate first-stage predictions. These are subsequently integrated through a residual meta-learning module to produce the final estimates. Evaluated on the LUCAS-2009 and AfSIS datasets, the framework consistently outperforms strong baselines in Coefficient of Determination (R^2) and Root Mean Square Error (RMSE), demonstrating its robustness and potential for high-throughput, multi-target soil property assessment.**

***Index Terms*—Deep Learning, Hyperspectral, Soil Properties, Transformer, CNN, Particle Swarm Optimization**

I. Introduction

Soil forms the foundation of terrestrial ecosystems and sustains human life by supporting agricultural productivity, regulating water cycles, and enabling nutrient cycling [1]. Its quality depends on a delicate balance of physicochemical properties, among which Calcium (Ca), pH, Phosphorus (P), sand content, and Soil Organic Carbon (SOC) play pivotal roles [2].

Accurate prediction of soil properties is critical for sustainable land management and environmental conservation. However, achieving high accuracy is challenging due to soil heterogeneity and complex interdependencies among properties [3]. Conventional laboratory methods, though precise, require extensive sampling and chemical analysis, making them costly and slow to scale [4]. In contrast, hyperspectral reflectance spectroscopy provides a rapid, non-destructive, and cost-effective alternative capable of simultaneously estimating multiple key soil properties from minimal samples, offering a highly promising solution for accurate nutrient profiling [5].

Several prior studies on LUCAS-2009 [6] and AfSIS [7] datasets, such as Cao et al. (2024) [8] and Zhao et al. (2023) [3], have achieved notable success in soil property prediction using deep learning. However, these methods often face limitations such as reliance on a single architecture type or fixed fusion of architectures. In contrast, this paper explores multiple fusion strategies, including independent, concatenated, and adaptively weighted combinations of Convolutional Neural Network (CNN) and Transformer branches. A similar hybrid approach has been used by Cao et al. (2024) [8], who have combined CNN and Transformer architectures by concatenating their feature embeddings. Such designs assume both branches contribute equally and overlook the possibility that the optimal integration of modalities may vary depending on dataset characteristics and target properties.

Other studies, such as Jin et al. (2023) [9] and Akande et al. (2022) [10], directly model all bands without feature selection, which can lead to overfitting and degraded predictive performance. While Principal Component Analysis (PCA) is widely used to reduce dimensionality, relying on it alone risks preserving variance from irrelevant or noisy bands, which can compromise accuracy. To address this, the paper first applies Particle Swarm Optimization (PSO) for feature selection, enabling the identification of the most informative spectral bands directly linked to the prediction targets, and then employs PCA on the reduced set.

To address these limitations, the contributions of this paper are as follows:

1) Developing Multi-Strategy Fused Transformer-CNN (MSF-TCNN), a novel Transformer-CNN fusion strategy to optimally integrate local and global spectral information for improved soil property prediction accuracy.

979-8-3315-4970-1/26/$31.00 © 2026 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on April 22,2026 at 15:00:50 UTC from IEEE Xplore. Restrictions apply.



INTERNATIONAL CONFERENCE ON AGRICULTURE-CENTRIC COMPUTATION
IIT Ropar, 11-13th May, 2023

CERTIFICATE OF PRESENTATION

This is to certify that

Shagun Jain

has successfully presented the paper titled

A review on applications of artificial intelligence for identifying soil nutrients

at the ICA-2023.

Neeraj Goel.

Dr. Neeraj Goel
Indian Institute of Technology, Ropar
Conference Chair

Dr. Mukesh Kumar Saini
Indian Institute of Technology, Ropar
Conference Chair

2023 First IEEE International Conference on

Advances in Electrical, Electronics and Computational Intelligence (ICAEECI 2023)

19 - 20, October 2023 | Tiruchengode, Tamil Nadu, India

CERTIFICATE

This certificate is presented to



Shagun Jain
Department of Software Engineering,
Delhi Technological University, New Delhi, India.

KCS 6007

for presenting the research paper entitled "A Hybrid Approach for Soil Nutrient Estimation using Multispectral Data in Belgium and Luxembourg" in the First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI 2023) held at K.S.R. College of Engineering, Tiruchengode, Tamil Nadu, India during 19 - 20, October 2023. The conference has been organized by the department of Electrical & Electronics Engineering, K.S.R. College of Engineering.

Dr. M.Ramasamy
Organizing Chair

Dr. S. Ramesh
Conference Chair

Dr.P.Senthilkumar
Patron

Organized by

K.S.R. COLLEGE OF ENGINEERING
(An autonomous institution, Accredited by NBA & NAAC)
Tiruchengode, Tamil Nadu, INDIA
Website: www.ksrce.ac.in

Partners

CMS Partner

Publication Partner



2ND INTERNATIONAL CONFRENCE ON AGRICULTURE-CENTRIC COMPUTATION

21ST-24TH MAY, 2024

Certificate of Presentation

This is to certify that

Shagun Jain

has successfully presented the paper titled

Comparative Analysis of Machine Learning based Soil pH prediction using Spectral Bands & Indices.

at the ICA-2024

Dr. Rajiv Ratn Shah
Conference Chair, IIIT Delhi

Dr. Mukesh Kumar Saini
Program Chair, IIT Ropar



You Choose, We Do It

St. JOSEPH'S COLLEGE OF ENGINEERING

(An Autonomous Institution)

OMR, Chennai - 119.

First International Conference on

FRONTIER TECHNOLOGIES AND SOLUTIONS - ICFTS 2025

CERTIFICATE

ICFTS 329

Peer Reviewed

Publication Partner

IEEE

Advancing Technology for Humanity

This certificate is presented to



Ms. Shagun Jain
Ph.D Scholar,
Department of Software Engineering
Delhi Technological University
Delhi-110042, India

for presenting the research paper entitled "**CNViS-Net: A Self-Attention Framework for Predicting Soil Nutrients from Hyperspectral Data**" in the IEEE sponsored First International Conference on Frontier Technologies and Solutions (ICFTS 2025) held at the Department of Electrical and Electronics Engineering, St. Joseph's College of Engineering, Chennai, Tamil Nadu, India, during 27 - 28, March 2025.

Lt. R. Sreekanth, M.E., (Ph.D.)
Organising Chair

Dr. Jayarama Pradeep, M.E., Ph.D.
Conference Chair

Dr. Vaddi Seshagiri Rao, M.E., Ph.D.
Principal

Phdthesis

ORIGINALITY REPORT

9%	6%	8%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Agricultural-Centric Computation", Springer Science and Business Media LLC, 2025 Publication	<1%
2	R.K. Naresh, Rajan Bhatt, Krishan K. Verma, Hari Om Katiyar, Arun Kumar, Mauro Wagner de Oliveira. "Artificial Intelligence in Regenerative Agriculture - Transforming Soil Health and Crop Productivity", Elite Publishing House, 2026 Publication	<1%
3	"Emerging Trends and Technologies on Intelligent Systems", Springer Science and Business Media LLC, 2025 Publication	<1%
4	Submitted to Global Banking Training Student Paper	<1%
5	ir.vistas.ac.in Internet Source	<1%
6	dtu.ac.in Internet Source	<1%
7	Submitted to Krea University Student Paper	<1%
8	ouci.dntb.gov.ua Internet Source	<1%
9	nomresearch.cn Internet Source	<1%
10	esdac.jrc.ec.europa.eu Internet Source	<1%
11	R.K. Gupta, R.K. Naresh, S.S. Walia, U.P. Singh, Dinesh Tomar, Ashok Kumar, Y.P. Singh, Raj Paul Sharma. "AI-Driven Smart Fertilization - Enhancing Crop-Water Productivity and Soil Health", CRC Press, 2026 Publication	<1%
12	Akshyat Shah, Shashi Sah, Shashvat Singhal, Shagun Jain, Divyashikha Sethi. "DSeP-xNet: A Feature Optimized Ensemble Framework for Hyperspectral-Based	<1%

SHAGUN JAIN

shagunjain191172@gmail.com

shagun-jain | shagun111997 | Shagun-Jain | Shagun-Jain

Delhi - 110034, India

SUMMARY

Focused on advancing research and academia, with a strong background in Artificial Intelligence, software development, and academic research. Currently serving as an Assistant Professor at Vivekananda Institute of Professional Studies–Technical Campus (VIPS-TC), with two years of prior industry experience as an SDE-II at Busy Infotech Pvt. Ltd. UGC NET qualified and pursuing a PhD at Delhi Technological University since August 2022 in Smart Agriculture and Artificial Intelligence. Research interests include soil science, remote sensing, satellite image processing, computer vision, and AI-driven solutions for agriculture. Dedicated to fostering innovation, mentoring students, and contributing to academic excellence and impactful research. Honoured with the Research Excellence Award 2025 by DTU.

EXPERIENCE

- **Vivekananda Institute of Professional Studies–Technical Campus (VIPS-TC) [🌐]** *Nov 2025 - Present*
  Assistant Professor Delhi, India
  - Teaching undergraduate and postgraduate courses in Computer Applications, with emphasis on programming, artificial intelligence, and core computing concepts.
  - Mentoring students in academic projects, research activities, and technical development.
  - Contributing to research, publications, and institutional research and innovation initiatives.
- **Delhi Technological University [🌐]** *Aug 2022 - Oct 2026*
  Research Scholar Delhi, India
  - Worked on satellite image processing and implemented neural networks for soil nutrient prediction in precision farming.
  - Guided B.Tech Computer Science Engineering students and co-authored conference papers.
- **IARI (Indian Agricultural Research Institute, Delhi) [🌐]** *April 2023 - July 2023*
  Summer Training Delhi, India
  - Worked on preprocessing of UAV-based hyperspectral imaging through ENVI and ArcGIS.
  - Implemented a methodology for the prediction of soil properties using hyperspectral data and machine learning algorithms.
- **BUSY INFOTECH PVT LTD. (Accounting Software Company) Acquired by IndiaMART [🌐]** *Oct 2020 - July 2022*
  SDE II - Development Delhi, India
  - Handled both Backend Services as well as Frontend Tier on the .Net framework.
  - Developed optimized SQL Queries to fetch data from a large database for business reports.

EDUCATION

Institution	Degree	Grade	Year
Delhi Technological University, Delhi	PhD in CS	9.0 SGPA	Aug 2022 - Oct 2026
Bharati Vidyapeeth's Institute of Computer Applications and Management, GGSIPU, Delhi	MCA	89%	June 2020
Vivekananda Institute of Professional Studies, GGSIPU, Delhi	BCA	93.2%	June 2018

PUBLICATIONS

J=JOURNAL

[J.1] **Shagun Jain**, Divyashikha Sethia & K.C. Tiwari, "A critical systematic review on spectral-based soil nutrient prediction using machine learning", Environmental Monitoring and Assessment 196, 699 (2024). SCIE, IF= 3.6, July 2024. https://doi.org/10.1007/s10661-024-12817-6

[J.2] **Shagun Jain**, Divyashikha Sethia & K.C. Tiwari, "Developing novel spectral indices for precise estimation of soil pH and organic carbon with hyperspectral data and machine learning", Environmental Monitoring and Assessment, 196(12), 1255, SCIE, IF=3.6, Nov 2024. https://doi.org/10.1007/s10661-024-13406-3

[J.3] **Shagun Jain**, Divyashikha Sethia & K.C. Tiwari, "CAT-MTLNet: A Novel Deep Multi-Target Framework for Hyperspectral Soil Properties Estimation Using Concrete Autoencoder and Transformer", IEEE Geoscience and Remote Sensing Letters, SCIE, IF=4.8, Publisher: IEEE. https://doi.org/10.1109/LGRS.2025.3635157

[J.4] **S. Jain**, D. Sethia & K.C. Tiwari, "TLM-Stack: A Deep Learning-based Novel Framework for Soil Nutrients Estimation Using Hyperspectral Data", Computers and Electronics in Agriculture, SCIE, Elsevier, IF=10.3, March 2026. https://doi.org/10.1016/j.compag.2025.111381

[J.5] **S. Jain et al.**, "ExGAPS:A Unified and Optimized Explainable AI-Driven Framework for Multivariate Hyperspectral Soil Property Estimation", Engineering Applications of Artificial Intelligence, SCIE, Elsevier, IF=9.0, April 2026. https://doi.org/10.1016/j.engappai.2026.114257

WORKSHOPS

- Participated in Youth Empowerment and Skills Workshop (YES!+) organized by "The Art of Living". *Feb 2016*
- Participated in Systematic Literature Review Workshop organized by Commacad. *Jan 2023*
- Attended Skill Enhancement Computer Workshop organized by Delhi Technological University. *Oct 2023*
- Participated in One Day Workshop on Leveraging the Role of Computer Vision on Agriculture Applications organized by International Rice Research Institute and Delhi Technological University. *Nov 2023*
- Participated in One Day International Workshop on Hyperspectral Remote Sensing conducted by IIT Tirupati Navishankar I-Hub Foundation. *May 2024*

Biography - Supervisor



Dr. Divyashikha Sethia has more than 5 years of rich industrial experience in telecom and networking areas in leading companies in India and the US, such as Cisco Systems , 2Wire, Force10 Networks, and Future Software (acquired by Aricent). She has more than 18 years of experience in research and academics. She brings rich industry experience to her teaching and research in academia. She leads the Mobile Computing Distributed System (MsDS) research lab, MIND Lab, under the Center of Excellence of Science of Happiness, and co-leads the Samsung lab and MIND lab at DTU. Her research interests are AI for healthcare, physiological signal-based cognitive healthcare and smart agriculture, Secure portable health record systems, Mobile computing, IoT, Cryptography, distributed systems, Networked multimedia and video streaming.

Biography - Co-Supervisor



Dr (Col) Kailash Chandra Tiwari is a professor of Geospatial Science and Technology. He has served in the Indian Army Corps of Engineers for 20 years. Dr Tiwari has

pursued PhD in Target Detection Using Optical and Microwave Remote Sensing, prior to this has a Masters in Engineering in CAD (Civil) and Post-Graduate Diploma in International Marketing. He has an extensive experience of teaching and research in remote sensing, military target detection and image processing.

Biography - Student



Shagun Jain is a Research Scholar at Delhi Technological University, Delhi, India. Her research interests include Artificial Intelligence, smart agriculture, remote sensing, satellite image processing, and soil nutrient prediction using machine learning and deep learning techniques. She has qualified the UGC-NET examination and received the DTU Research Excellence Award in 2025 for her research contributions. She has two years of industry experience as a Software Development Engineer and one year of teaching experience. Her work focuses on developing AI-driven solutions for agricultural applications and precision farming.