الجمهورية الجزائرية الديمقراطية الشعبية

People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي

Ministry of Higher Education and Scientific Research

جامعة غرداية

University of Ghardaia

كلية العلوم والتكنولوجيا

Faculty of Science and Technology

قسم الرياضيات والإعلام الآلي

Department of Mathematics and Computer Science

مخبر الرياضيات والعلوم التطبيقية

Mathematics and Applied Sciences Laboratory

Registration n°:
...../...../...../...../.....

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

Master

Domain: Mathematics and Computer Science

Field: Computer Science

Specialty: Intelligent Systems for Knowledge Extraction

Topic

Soil Quality Prediction using Machine Learning

Presented by:

MOKDAD Meriem & TALEB AHMED Abdelmalek

Publicly defended on June 30, 2025

Jury members:

Mr. Bellaouar Slimane	MCA	Univ. Ghardaia	President
Mr. Mahdjoub Youcef	MAA	Univ. Ghardaia	Examiner
Mr. Oulad Naoui Slimane	MCB	Univ. Ghardaia	Supervisor
Mrs. Ben abderrahmane Habiba	PhD Student	Univ. Laghouat	Co-Supervisor

Academic Year: 2024/2025

Acknowledgment

This master's thesis reflects the result of the assistance and teamwork of many individuals to whom we are greatly indebted.

We would like, first and foremost, to thank our supervisor, Dr. Slimane Oulad-Naoui, Associate Professor at the University of Ghardaia, for his valuable help, permanent support, and unremitting patience throughout the elaboration of this work, in spite of the many challenges faced along the way.

We are also grateful to our co-supervisor, Habiba Ben-Abderrahmane, PhD student at the University of Laghouat, whose reflective guidance and unique mentoring style were essential in helping us define and address our research problem.

We also wish to express our sincere gratitude to the great members of the jury for their interest in our work and for the honor of having this thesis assessed.

Finally, we thank the Applied Mathematics Laboratory of the University of Ghardaia, under the leadership of Dr. Slimane Bellaouar, for granting us access to required resources and for their hospitality.

إهداء

أولاً وأبداً، الحمد لله الذي بنعمته تتم الصالحات، وبفضله تتحقق الأمنيات. نحمده ونشكره على ما وهبنا من صبرٍ وقوة، وما يسّر لنا من سُبُل العلم والعمل، فله الحمد كما ينبغي لجلال وجهه وعظيم سلطانه.

في لحظةٍ امتزجت فيها مشاعر الفخر بالامتنان، وبين طيّات هذا العمل الذي يُتوَّج سنواتٍ من الجهد والسهر، أقف شاكرًا وممتنًا لكلّ من كان له الأثر في رسم معالم هذا الإنجاز.

إلى والديّ العزيزين ، نبض قلبي وسرّ عطائي، أنتما السند الذي لم يتخلّ عني، والدعاء الذي رافقني في كلّ خطوة، والملجأ الذي أعود إليه في لحظات الشك. لولا تضحياتكما، لما كنت ما أنا عليه اليوم. فشكرًا لكما بعدد أنفاسي، وبعمق السماء.

إلى إخوتي، رفاق الدرب، من شاركوني لحظات التعب والتوتّر، وفرحة التقدّم والإنجاز، فلكم مكانة لا يطالها النسيان في قلبي. دعمكم ومساندتكم كانت الزاد الذي أعانني على الاستمرار.

إلى نفسي...

إلى تلك الروح التي لم تستسلم، رغم العثرات، والإرهاق، والخذلان أحيانًا... إليكِ يا أنا، يا من وقفتِ في وجه الخوف، وتجاوزتِ التعب، وحملتِ الحلم رغم ثقله.

أُهديكِ هذا الإنجاز، لأنه من صنعكِ، ومن نبضكِ، ومن لياليكِ البيضاء. أُهديه لكِ فخرًا، وعرفانًا، وامتنانًا، لأنكِ لم تتراجعي... لأنكِ أكملتِ الطريق حتى نهايته، برأسٍ مرفوع، وقلبٍ مفعمٍ بالإصرار.

دمتِ قويّة، طموحة، مخلصة، وصادقة مع ذاتكِ، لا تنكسرين مهما اشتدّ الطريق.

وإلى زميل الدرب، "طالب أحمد عبد المالك" أُهديك هذا النجاح الذي حققناه سويًا بكل امتنان وتقدير... كنتَ حاضرًا بالصبر، والدعم، والرفقة الطيّبة في كل خطوة... بأفكارك، واجتهادك، وروحك المتعاونة، كنتَ عونًا لا يُقدّر... شكرًا لك من القلب، ولك مني كل التمنيات بالمزيد من التوفيق والنجاح.

مقداد مريم

إهداء

في لحظةٍ تختلط فيها مشاعر الفخر بالشكر، وبين سطور هذا العمل الذي يُكلّل مسيرة أعوامٍ من المثابرة والاجتهاد، أقف ممتنًا لكل من أسهم في رسم ملامح هذا الإنجاز.

إلى والديّ العزيزين، أنحني تقديرًا واحترامًا أمام عطائكما الذي لا يُقارن، وصبركما الذي لا يُحد، ومحبتكما التي أحاطتني منذ نعومة أظافري وحتى هذه اللحظة.

لقد كنتما النور الذي أضاء دربي، والدافع الذي بعث في قلبي الأمل حين خفت، والقوة التي تسندني عندما أثقلتني المسؤولية. نجاحي هذا هو ثمرة تعبكما، وقطاف سنين سهر وصبر ودعاء. لكما مني كل الشكر، وكل الحب، وكل الإهداء. **وإلى إخوتي وأخواتي،** أنتم السند، والكتف، والدعم الذي لا يتغير مهما تغيّرت الظروف.

لكل كلمة تشجيع، ولكل لحظة اهتمام، ولكل دعوة صادقة، أقول شكرًا من القلب.

لكم جميعًا، أهدي هذا العمل، محبةً وامتنانًا ووفاءً... ودعائي أن أكون دائمًا عند حسن ظنكم، فخورين بي كما أنا فخور بانتمائي إليكم.

إلى نفسي...

إليكِ أيتها الروح الصامدة، التي وقفت شامخة رغم كل العثرات، والتي قاومت التعب والخذلان، ولم تتخلَّ عن الحلم يومًا... أُهديكِ هذا الإنجاز بكل فخر، لأنه ثمرة صبركِ، وسهر لياليكِ، وإصراركِ على الاستمرار رغم كل شيء.

وإلى زميلة الدرب، "مقداد مريم" أهديكِ هذا النجاح بكل محبة وتقدير... كنتِ الرفيقة الصبورة، والداعمة المخلصة، والحاضرة بروحكِ الإيجابية وأفكاركِ النيّرة في كل مرحلة. باجتهادكِ وتعاونكِ، كان الطريق أيسر، والرحلة أغنى. شكرًا من القلب، ودعواتي لكِ بمزيد من التميز والتوفيق في قادم الأيام.

وفي الختام، الحمد لله الذي بنعمته تتمّ الصالحات، ونسأله أن يكون هذا العمل خطوة مباركة في طريق العلم والعطاء.

طالب أحمد عبد المالك

ملخص

تلعب عملية التنبؤ بجودة التربة (SQP) دورًا حاسمًا في الزراعة، وإدارة البيئة، والهندسة المدنية. تُعد الطرق التقليدية في التقييم، مثل التحاليل المخبرية والمسح الميداني، مكلفة وتستغرق وقتًا طويلاً، كما أنها محدودة من حيث التغطية الجغرافية. يهدف هذا العمل إلى تطوير نظام ذكي للتنبؤ بجودة التربة واقتراح النباتات المناسبة باستخدام تقنيات التعلم الآلي والنمذجة الجغرافية. ولتحقيق هذا الهدف، تم تنفيذ تجربتين رئيسيتين. في التجربة الأولى، تم استخدام أربعة نماذج: RBFN، LightGBM، XGBoost، وDNN على بيانات من منصة SoilGrids، تتضمن خصائص فيزيائية وكيميائية للتربة. وقد حقق نموذج XGBoost أفضل أداء بدقة $R^2 = 0.98$، مما يؤكد فعاليته في مهام التنبؤ بجودة التربة. أما التجربة الثانية، فاعتمدت على بنية تنبؤ من مرحلتين: في المرحلة الأولى، تم تدريب 63 نموذجًا انحداريًا مستقلًا للتنبؤ بخصائص التربة والبيئة انطلاقًا من الإحداثيات الجغرافية. ثم استُخدمت هذه التنبؤات في نموذج Random Forest لحساب مؤشر جودة التربة (SQI). وفي المرحلة الثانية، تم اعتماد طريقة تعتمد على Cosine Similarity لمقارنة الظروف المتوقعة للموقع مع متطلبات النباتات المثلى واقتراح الأنواع الأنسب. تم نشر النظام بالكامل على شكل تطبيق ويب تفاعلي، يتيح للمستخدمين الاستعلام عن خرائط SQI في الزمن الحقيقي، والحصول على توصيات زراعية مخصصة. من بين التحسينات المستقبلية: إدماج بيانات محلية أدق، خاصة من الجزائر، وتوسيع التغطية الجغرافية للنظام.

كلمات مفتاحية: التنبؤ بجودة التربة، التعلم الآلي، النمذجة الجغرافية المكانية، نظام اقتراح النباتات.

Abstract

Soil quality prediction (SQP) plays a crucial role in agriculture, environmental management, and civil engineering. Traditional assessment methods, such as laboratory analyses and field surveys, are often time-consuming, costly, and limited in spatial coverage. This work aims to develop an intelligent system for predicting soil quality and recommending suitable crops using machine learning and geospatial data. To achieve this, two key experiments were conducted. In the first experiment, four models (RBFN, LightGBM, XGBoost, and DNN) were applied to SoilGrids data, including physical and chemical characteristics of the soil. The XGBoost model achieved the best performance $R^2 = 0.98$, reasserting its suitability for SQP tasks. The second experiment used a two-stage prediction architecture. The first stage trained 36 separate regressors to predict soil and environmental conditions from geolocation data. These predictions were then used in a Random Forest model to estimate the Soil Quality Index (SQI). The second stage employed a cosine similarity-based method to recommend the most suitable plant species based on the predicted site conditions. The entire system was deployed as an interactive web application, where users can query real-time SQI maps and receive personalized crop recommendations. Future enhancements include incorporating more localized data, particularly from Algeria, and expanding the spatial coverage of the system.

Keywords: Soil Quality Prediction, Machine Learning, Geospatial Modeling, Plant Suggestion System.

Résumé

La prédiction de la qualité du sol (SQP) joue un rôle crucial dans l'agriculture, la gestion de l'environnement et le génie civil. Les méthodes traditionnelles d'évaluation, telles que les analyses en laboratoire et les enquêtes de terrain, sont souvent longues, coûteuses et limitées en couverture spatiale. Ce travail vise à développer un système intelligent capable de prédire la qualité du sol et de recommander des cultures adaptées, en utilisant l'apprentissage automatique et les données géospatiales. Pour cela, deux expériences principales ont été menées. Dans la première expérience, quatre modèles ont été testés : RBFN, LightGBM, XGBoost et DNN, en utilisant les données de SoilGrids contenant des caractéristiques physiques et chimiques du sol. Le modèle XGBoost a obtenu les meilleurs résultats avec un score de $R^2 = 0.98$, confirmant son efficacité pour les tâches de SQP. La deuxième expérience repose sur une architecture de prédiction en deux étapes. La première consiste à entraîner 36 régressions indépendantes pour prédire les caractéristiques du sol et de l'environnement à partir des coordonnées géographiques. Ces prédictions sont ensuite utilisées dans un modèle Random Forest pour estimer l'indice de qualité du sol (SQI). La deuxième étape applique une méthode basée sur la Cosine Similarity afin de recommander les plantes les plus adaptées en fonction des conditions environnementales prévues. Le système final est déployé sous forme d'une application web interactive permettant d'accéder à des cartes SQI en temps réel ainsi qu'à des recommandations personnalisées de cultures. Les améliorations futures incluent l'intégration de données locales, notamment en provenance d'Algérie, et l'élargissement de la couverture spatiale du système.

Mots clés: Prédiction de la qualité des sols, apprentissage automatique, modélisation géospatiale, système de suggestion de plantes.

Contents

List of Figures

List of Tables

List of Acronyms

A-SQI	Additive Soil Quality Index
CNN	Convolutional Neural Network
DNN	Deep Neural Network
GIS	Geographical Information System
HLSF	Homothetic Linear Scoring Function
IDW	Inverse Distance Weighted
LightGBM	Light Gradient Boosting Machine
ML	Machine Learning
PCA	Principal Component Analysis
R^2	Coefficient of Determination
RBFN	Radial Basis Function Network
RF	Random Forest
RMSE	Root Mean Square Error
SQ	Soil Quality
SQI	Soil Quality Index
W-SQI	Weighted Soil Quality Index
XGBoost	eXtreme Gradient Boosting

Introduction

Soil quality is not merely at the heart of farm performance but also of environmental sustainability. Good soil quality assessment guides decision-making in crop selection, fertilizer use, and sustainable land management. Conventional approaches, however, are based on extensive field sampling with subsequent laboratory analysis that is resource-consuming, time-consuming, and spatially limited.

With the recent expansion in the availability of geospatial and environmental data, an increasing emerging application of data-intensive approaches has appeared. Among them, machine learning (ML) methods have been promising in providing automated Soil Quality Prediction (SQP). Despite this, practical solutions that integrate ML with spatial modeling for SQP are rare, especially localized, scalable, and simple-to-use decision-making tools.

This study investigates the application of four machine learning models—Radial Basis Function Networks (RBFN), LightGBM, XGBoost, and Deep Neural Networks (DNN)—to predict soil quality from geospatial information. Nine major soil parameters—pH, nitrogen (N), phosphorus (P), potassium (K), clay, silt, sand, bulk density, and organic carbon—are covered in the dataset that is derived from SoilGrids. The models are compared to identify the best performing in the prediction of soil traits at a broad range of locations.

Based on this, a two-stage artificial intelligence pipeline is suggested. The highest performing model in stage one predicts environmental and soil features from geographic coordinates. The predictions are used to estimate the Soil Quality Index (SQI) through a Random Forest Regressor. In stage two, a recommendation system employs cosine similarity to pair site conditions with plant growth requirements and supply appropriate crop recommendations.

The system is designed as a web application offering soil analysis and plant recommendations. The aim of this work is to provide an extensible and adaptable solution to assist farmers, agronomists, and decision-makers in improving farm planning and land management.

The remainder of the thesis is structured as follows: Chapter 1 introduces the background ideas on soil properties, geospatial modeling, and algorithms utilized. Chapter 2 is an overview of prior relevant work in soil prediction and plant-environment compatibility. Chapter 3 talks about experimental method, system design, and testing and then current limitations and future trends.

CHAPTER 1

BASIC CONCEPTS

1.1 Introduction

This chapter is dedicated to the background of soil quality, Geospatial Information System (GIS) as an important tool of mapping, MLs techniques used to study the soil quality.

1.2 Soil Quality

The quality of soil plays a fundamental role in multiple aspects of life, particularly in agriculture, environmental sustainability, and land management.Sumathi et al. (2023). As originally proposed in the early 1990s, soil quality refers to the ability of a soil to perform its functions. More precisely, according to the USDA (1994)[1], soil quality is defined as the capacity of a specific type of soil, within its natural or managed ecosystem boundaries, to sustain plant and animal productivity, improve or preserve air and water quality, and support human health and ecosystems. Soil fulfills five vital functions: it sustains plant and animal life, regulates water flow, filters and breaks down contaminants, recycles nutrients, and provides structural support. Sepehya et al. (2024)

1.2.1 Soil Indicators

Over the past three decades, the concept of soil quality (SQ) has evolved to reflect the soil's ability to perform essential ecological functions. It is determined by the integration of physical, chemical, and biological characteristics, though it cannot be directly measured. Instead, SQ is inferred through a range of indicators that respond to environmental changes and land management practices. These include commonly assessed properties which are illustrated in Figure 1.1 such as pH, electrical conductivity (EC), cation exchange capacity (CEC), soil organic matter

[1] USDA: United States Department of Agriculture

(SOM), and nutrient levels (N, P, K), all of which influence fertility, structure, porosity, and ecosystem services.

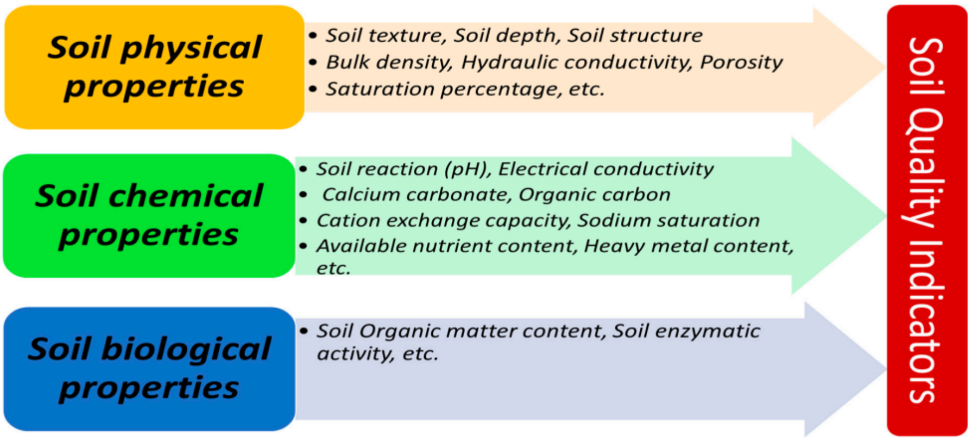

Soil physical properties
• Soil texture, Soil depth, Soil structure
• Bulk density, Hydraulic conductivity, Porosity
• Saturation percentage, etc.
Soil chemical properties
• Soil reaction (pH), Electrical conductivity
• Calcium carbonate, Organic carbon
• Cation exchange capacity, Sodium saturation
• Available nutrient content, Heavy metal content, etc.
Soil biological properties
• Soil Organic matter content, Soil enzymatic activity, etc.
Soil Quality Indicators

Figure 1.1: Different Soil quality indicators El Behairy et al. (2024a)

SOM is also highlighted specifically as a key indicator, given its importance in soil aggregation, nutrient cycling, carbon sequestration, and general fertility. Literature indicates that the use of organic amendments to raise SOM improves soil health and crop yields. Phosphorus and nitrogen are also important, with P frequently being the second most limiting nutrient after N. Soil pH plays a central role in nutrient availability, SOM decomposition, and microbial activity, to be touted as the ”master variable.Calcium carbonate (CaCO) has a beneficial effect through its ability to increase water-holding capacity and reduce hydraulic conductivity. Soil texture also exerts a powerful influence on water retention, aeration, and root development. Coarse deep soils can have poor water retention potential, affecting productivity. Lastly, while all types of indicators are important, biological indicators are found to be more sensitive and immediate in their response to environmental change, and hence particularly valuable for soil quality evaluation and guiding management decisions. They offer a missing link between soil condition and crop performance for more sustainable land use systems.El Behairy et al. (2024a)

1.2.2 Soil Quality Index

In order to evaluate soil quality, a metric named **Soil Quality Index** is computed. Several methods are proposed, in Damiba et al. (2024) two methodologies were utilized to compute the SQI: the Additive Soil Quality Index (A-SQI) and the Weighted Soil Quality Index (W-SQI).

The **Additive Soil Quality Index (A-SQI)** approach involves calculating the arithmetic mean of the normalized scores of all selected soil indicators. Normalization is achieved using a Homothetic Linear Scoring Function (HLSF) that standardizes indicator values between 0 and 1. The formula for A-SQI is expressed as:

$$\text{A-SQI} = \frac{1}{n}\sum_{i=1}^{n} S_i \tag{1}$$

where S_i represents the normalized score of the i-th soil indicator, and n is the total number of the considered indicators. This method treats all indicators equally without assigning specific importance, making it simple and broadly applicable when the relative impact of indicators is unknown.

In contrast, the **Weighted Soil Quality Index (W-SQI)** refines the A-SQI by incorporating the relative importance of each soil indicator. Principal Component Analysis (PCA) is first applied to identify key indicators and assign weights based on the proportion of variance explained by each principal component. Each normalized score is multiplied by its corresponding weight, and the weighted scores are summed to obtain the final index. The W-SQI is computed as:

$$\text{W-SQI} = \sum_{i=1}^{m} \omega_i \times S_i \tag{2}$$

where ω_i denotes the weight assigned to the i-th soil indicator derived from PCA. This method provides a more sensitive and precise evaluation of soil quality, particularly when certain indicators have a greater influence on soil functionality.

Additionally, the sensitivity of each SQI method was evaluated to determine their responsiveness to differences in soil conditions. The sensitivity index (S) was calculated as follows:

$$\text{Sensitivity } (S) = \frac{\text{SQI}_{\text{max}}}{\text{SQI}_{\text{min}}} \tag{3}$$

where SQI_{max} and SQI_{min} represent the maximum and minimum SQI values among all study sites, respectively. This metric is useful for assessing the ability of the SQI methods to distinguish between different soil quality levels.

A summary comparison between A-SQI and W-SQI is presented in Table 1.1.

Table 1.1: Comparison of A-SQI and W-SQI methodologies

Criterion	**A-SQI**	**W-SQI**
Indicator selection	All indicators are equally included	Selected indicators via PCA
Weight assignment	No weights assigned (equal importance)	Weights ω_i based on explained variance
Formula	$\text{A-SQI} = \frac{1}{n}\sum_{i=1}^{n} S_i$	$\text{W-SQI} = \sum_{i=1}^{m} \omega_i \times S_i$
Sensitivity to soil variability	Lower	Higher

One of our goals is to map the soil quality index, therefore, we present in Section 1.3 the existing mapping techniques.

1.3 Geographical Information Systems (GIS)

Geographical Information Systems (GIS) are computer systems based on hardware, software, and georeferenced data that can be used to collect, store, manage, process, analyse, and visualize both spatial and non-spatial information representing real-world geographic phenomena . Georeferenced data refers to any data that are linked to a location on the Earth's surface through the use of a geographic or projected coordinate system. GIS data are digital objects which represent real-world entities and are defined by: their geometric properties (spatial location), their attributes (characteristics associated with each object), and their topology (definition of how entities are related to others in space). In other words, data provide means to locate them in space and can be overlaid, calculated, manipulated, visualised and analysed along with other data layers that use the same coordinate system. Each entity in the real world is represented by a data layer with geometric and topologic properties and an associated set of attributes, in the form of a table, which define the characteristics of that entity. GIS facilitates the analysis of spatial relationships within datasets based on the topological properties within the data. Topology refers to the inter connectivity and interrelated properties between data and defines and describes how spatial objects relate to their neighbors in space. Rogers et al. (2024)

1.3.1 Geo-Spatial data analysis techniques

One key concept in map visualization is **spatial interpolation** which is defined as using points with known values -sample-, the estimation of surface values at unsampled points -population-"(Chang, 2019).

Geo-Spatial interpolation techniques can be classified in various manners(Chang, 2019), one way is the *use points* : if all points are used then it's called **global interpolation** such as, *Trend Surface Models* and *regression*; Otherwise, it is called **local interpolation** using a part of known points. we cite : *Thiessen Polygons*, *Density Estimation*, *Inverse Distance Weighted (IDW)* interpolation and *Thin-Plate Splines.*

Another way is the *exactitude* of interpolation, where methods are grouped into exact and inexact interpolation, the first, predicts a value at the point location that is the same as its known value while the second predicts a value at the point location that differs from its known value.

The third way where spatial interpolation methods may be *deterministic or stochastic.* The deterministic interpolation methods provide no assessment of errors with predicted values, whereas the stochastic interpolation methods, considers the presence of some randomness in its variable and offers assessment of prediction errors with estimated variances. All the previously local methods, except kriging are deterministic. The kriging is a stochastic local method.

The previous methods applied mathematical models where the deterministic methods have no ground truth(Korstanje, 2023) and the stochastic are expansive in computation(Dramsch, 2020)(the computation of the semi-variogram for kriging is very expansive), it is necessary to search for other techniques that do the same tasks but with more efficiency. Thereafter why the machine learning techniques are used.

1.3.2 Machine Learning techniques for geospatial analysis

The goal of Machine learning (ML) is the creation of *models* that can learn from data (Flach, 2012), also ML provides a set of algorithms that can working with any type of data: labeled or unlabeled; unstructured, structured or semi-structured; sequential or unsequential; to solve different problems: classification (ANN, SVM, KNN, CNN...), regression, clustering (k-means, HAC,DBSCAN...), dimensionality reducing (PCA, FCA) sequential analysis (RNN, LSTM...).

Over the last 70 years (Dramsch, 2020) give an overview of the use of ML in spatial data analysis, or named Geo-science. The first use of Artificial Neural Networks in geographic was in 1980s, in seismic deconvolution with Hopfield neural network. while a Support Vector Machines (SVMs) were utilized for land usage classification using remote sensing early on 1999. However, Random Forests, faced a delay in achieving wider acceptance, primarily because it's name was not introduced until the year 2001.

For spatial interpolation, a new architecture of ANN, named FBRN, is used, where the activation function of the hidden layers is the "Radial Base Function" (RBF) proposed for the first time in 1987 by Powell, M.J.D(Lin and Chen, 2004), to solve the real multivariate interpolation problem. A Convolutional Neural Network (CNN) was applied for the first time on image seismic interpretation in 2017 by A.U. Waldeland and A. Solberg(Dramsch, 2020).

By applying ML sequential models Recurrent Neural Networks (RNN) and Long Short-Term Memories (LSTM), the spatial data analysis has a great development exploring different existing data in the field. (Dramsch, 2020) cited some applications. In 2017, researchers exploring unstructured text documents to extract the existing geological relation. Another application was seismological event classification of volcanic activity, multifactor landslide displacement prediction, sedimentological sequences modeling and prediction of petrophysical properties from seismic data.

Generative adversarial networks (GAN) were applied since 2017(Dramsch, 2020) to generate samples from data by several researchers in automatic seismic interpretation field.

1.4 Machine Learning (ML)

Machine learning involves developing algorithms that use data to attain domain expertise and facilitate decision-making autonomously. Machine learning algorithms differ from traditional programming in that they are data-oriented, making and learning from decisions based on what they are given. Machine learning is typically categorized into four general types: supervised, unsupervised, semi-supervised, and reinforcement learning—each having its own characteristics and applications.Guerraoui et al. (2024)

With the rapid development of machine learning technology, as a regression problem that helps people to find the law from the massive data to achieve the prediction effect, more and more people pay attention. Data prediction has become an important part of people's daily life. Currently, the technology is widely used in many fields such as weather forecasting, medical diagnosis and financial forecasting. Therefore, the research of machine learning algorithms in regression problems is a research hotspot in the field of machine learning in recent years. Huang et al. (2020)

Regression is the procedure of typical association between two or greater than two variables of interest concerning original elements of the data-set. It also command to launch the behavior of the association in the middle of variables on interest that are describing the practical association between the variables and thus afford an instrument for prediction or forecasting. It is a method of analysis and recognizes the relationship between two or greater than two variables of interest. The method that is adapted to execute regression analysis helps to realize whichever aspects are significant, whichever aspects fail to notice and in what way an individual promoting one and all. Regression castoff for prediction and forecasting. Regression is a subset of supervised machine learning techniques to predict the pattern of data.Kumar and Bhatnagar (2022)

1.4.1 Random Forest

Random Forest (RF, also called standard RF or BreimanRF) is an ensemble learning algorithm that makes classification or regression predictions by taking the majority vote or average of the results of multiple decision trees. Due to its simple and easy-to-understand nature, rapid training, and good performance, it is widely used in many fields, such as data mining, computer vision, ecology, and bioinformatics.LeCun et al. (2015) Random forests are a combination of tree predictors, as illustrated in Figure 1.2, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges almost surely. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them.Breiman (2001)

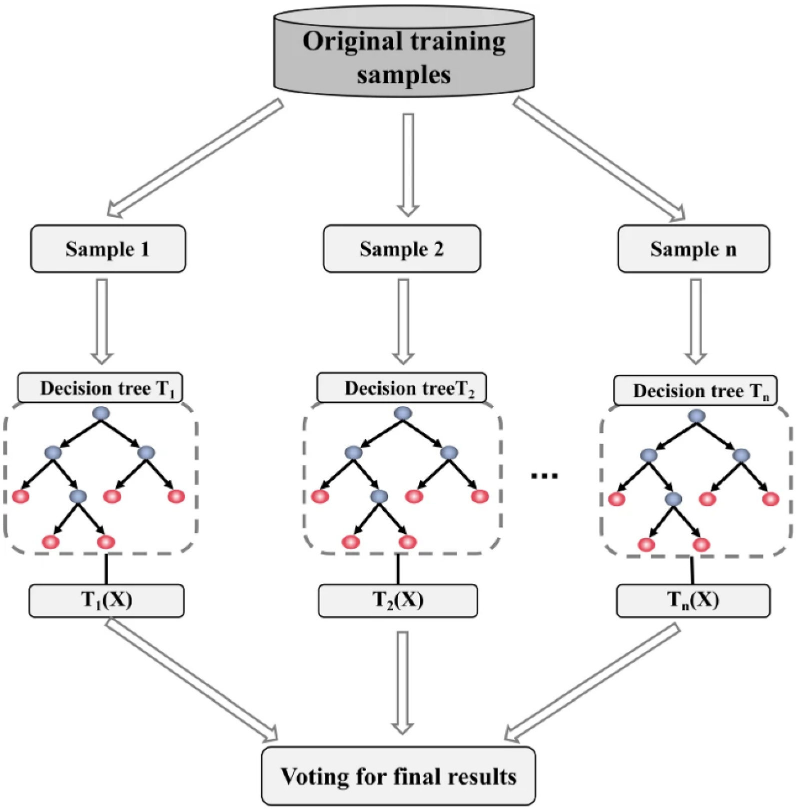


Figure 1.2: General architecture of a Random Forest
Yang et al. (2025)

1.4.2 XGBoost (Extreme Gradient Boosting)

The XGBoost (eXtreme Gradient Boosting) algorithm is a boosting-based ensemble learning method that accomplishes learning by constructing and integrating multiple weak learners. Its core principle involves iteratively adding different trees to the model, allowing it to evolve through feature splitting. Each newly added tree learns a new function, effectively fitting the residuals of the previous prediction. Ultimately, the final predicted value of a sample is obtained by summing the contributions of all trees in the model.Su et al. (2023)

1.4.3 Radial Basis Function Network (RBFN)

The radial basis function (RBF) neural network is usually employed due to its advantages such as its straightforward structure, higher estimation features, and a rapid training process. The RBF network is a potent feed forward neural network structure. A schematic diagram of a conventional RBF model is illustrated in figure 1.3.

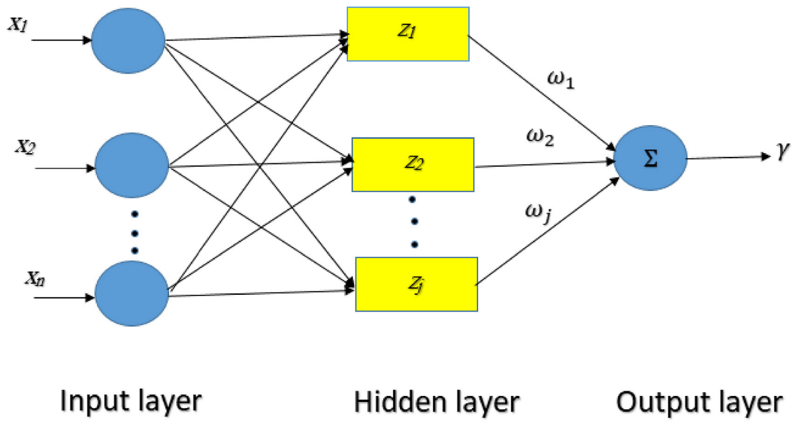


Figure 1.3: Structure of RBF model

The processing algorithm of the RBF model consists of three layers: an input, a hidden, and an output, respectively. All of the nodes in the layers are completely linked to the former layer. The inputs are allocated to each node in the input layer and then straightly delivered to the hidden layer. At last, weighted connections are used to transfer the data to the output layer. The significant stage of this model is the hidden layer where the RBF is applied as the activation function to produce the vector distance multiplied by the associated bias. Ramezanizadeh et al. (2019)

1.4.4 Deep Neural Networks (DNNs)

A **Deep Neural Network (DNN)** is one type of artificial neural network with several hidden layers between the input layer and the output layer. The DNNs are able to learn intricate concepts by constructing more abstract concepts out of easier features.Assert that the depth of the networks, their non-linearity, and the transformations applied at every layer make them extremely adept at grasping intricate relationships among large datasets.

DNNs contain an input layer that takes in raw data, one or more hidden layers that successively apply learned weights and activation functions to the data, and an output layer that generates predictions, which are typically task-specialized (e.g., softmax for classification). Adding more hidden layers enables the network to learn higher-level and abstract features, a fundamental concept of deep learning.

For alleviating typical problems such as vanishing gradients and overfitting, a number of solutions have been proposed. Some include using activation functions such as ReLU for preventing the gradient flow from being interrupted, and regularization techniques such as dropout, L2 regularization, and early stopping for improving generalization. Batch normalization is also typical for training stabilization and acceleration.

DNNs have been used with tremendous success in a wide range of applications, including computer vision, speech recognition, natural language processing, time series forecasting, and reinforcement learning. They are so versatile because they can learn features from the data itself, which reduces the need for feature engineering.Bishop and Bishop (2024)

1.5 Conclusion

In chapter one, we have had a detailed discussion of the fundamental concepts that form the foundation of our study of soil quality mapping. We have addressed the principles of soil quality evaluation and its major indicators as well as the quantitative measurement techniques such as the A-SQI and W-SQI methods. Furthermore, we have covered the use of GIS in spatial data analysis and talked about advanced computation techniques such as Random Forest, XGBoost, RBFN, and Deep Neural Networks. Having established this solid theoretical base, we are now poised to move forward into the challenging application of soil quality mapping techniques in the following chapters. We will explore the systematic evolution of these predictive models, assess pertinent research contributions, and investigate the technological innovations that have spurred progress in this area. This integrated philosophy will be our guiding principal as we strive to push the state of the art in soil quality evaluation and mapping technologies.

CHAPTER 2

STATE OF THE ART

2.1 Introduction

This chapter explores the principal work, and the results achieved in the past years on soil quality prediction. We divide our review into two sections, which are ML-based methods, and deep learning-based methods.

2.2 Traditional Methods

Brady and Weil (2016) emphasize that traditional soil testing involves a set of standardized laboratory tests developed over decades to determine important soil characteristics. These include:

- **Mechanical analysis** to determine the soil texture (sand, silt, clay fractions) using tests such as the hydrometer or pipette method.
- **Chemical analysis** for major nutrients such as nitrogen (via Kjeldahl digestion), phosphorus (via Bray or Olsen extraction), and potassium (using flame photometry or atomic absorption spectroscopy).
- **Soil pH** from water or salt solution (e.g., 1:1 soil-water solution).
- **Organic matter content** obtained from loss-on-ignition or Walkley-Black analysis.

They note that while these methods provide strong, benchmark data, they are spatially limited, require specialized laboratory equipment, and tend to lag due to transportation of samples and laboratory processing.

> “Although these tests are essential for understanding soil fertility, their applicability is often constrained when real-time or large-area assessments are needed.”

— Brady & Weil, 2016

While scientifically rigorous and standardized, they are time-consuming, labor-intensive, and costly. Furthermore, they have limited spatial generalizability because each soil sample represents only a small area. This has motivated growing interest in more scalable, data-driven approaches that complement or replace traditional testing, especially for extensive-scale environmental and agricultural planning.

2.3 Machine Learning methods

Du et al. (2020) investigates the structural characteristics of cutting notches in tea stalks using advanced X-ray micro-computed tomography technology to optimize harvesting equipment design for improved tea production efficiency. Researchers systematically analyzed third inter-node samples from Zhongcha 108 tea variety collected from Maichun tea farm in Zhenjiang, China, under precisely controlled cutting conditions with varying depths (0.7, 1.5, 2.3 mm) and cutting edge angles (30°, 35°, 40°). The experimental design employed a texture tester (Stable Micro Systems TA-XT2i) operating at 1.0 mm/s test speed to ensure consistent cutting parameters. Micro-CT scanning was performed using a Scanco Medical AG micro-CT 100 system at 45 kV and 88 A, generating 200 high-resolution slice images at 1024×1024 pixel resolution with 4 m maximum resolution capability. Image processing utilized sophisticated grey-scale histogram analysis and bimodal segmentation methods with optimal threshold values of 12 for data extraction, followed by 3D reconstruction and volume rendering using AVIZO software for comprehensive structural analysis.

Two critical quantitative metrics were established and validated: Maximum Cross-sectional Area Ratio of Cutting Notch (MCSARCN) and Volume Ratio of Cutting Notch (VRCN), both demonstrating proportional increases with cutting depth across all tested conditions. Specifically, MCSARCN values increased from 4.89% to 9.47% for 30° angle, 8.51% to 22.83% for 35° angle, and 4.30% to 16.87% for 40° angle as cutting depth progressed from 0.7 to 2.3 mm. Similarly, VRCN measurements showed increases from 1.59% to 2.13%, 2.98% to 5.76%, and 3.04% to 5.01% respectively. The 35° cutting edge angle demonstrated optimal performance characteristics, producing maximum VRCN values and establishing the most efficient cutting parameters. Conversely, when cutting depth was maintained constant at 1.5 mm, increased cutting edge angles resulted in higher cutting forces (14.20 N to 15.95 N) but paradoxically decreased VRCN values, revealing complex biomechanical interactions. The non-destructive imaging revealed intricate internal deformation patterns including inward shrinkage of stalk epidermis, irregular notch formation, and tissue compression effects. Notably, actual cutting notch depths consistently measured less than applied cutting depths due to the rheological properties and elastic recovery characteristics of tea stalk cellulose structure. This pioneering application of micro-CT technology to plant cutting mechanics provides quantitative foundations for evidence-based agricultural equipment optimization and represents a significant advancement in precision agriculture methodology.

Silvero et al. (2021) investigates the influence of spatial, spectral, and temporal resolutions of satellite images on predicting soil properties and their application to soil classification and management. The research was conducted over a 182-hectare area in southeastern Brazil, utilizing data from three different satellites: PlanetScope, Sentinel-2 MSI, and Landsat-8 OLI. A total of 120 topsoil samples were collected from the study area and analyzed for several key soil properties, including clay content, sand content, organic matter (OM), iron content (Fe_2O_3), and soil color characteristics (hue, value, and chroma). Multi-temporal synthetic soil images (SYSI), as well as single-date images, were employed as predictors in Cubist regression models to estimate the soil properties. The study evaluated various combinations of spectral bands, including visible to near-infrared (vis-NIR) and visible to near-infrared plus shortwave infrared (vis-NIR-SWIR) bands, using 10-fold cross-validation to assess the model performance. The results indicated that multi-temporal Sentinel-2 MSI images, particularly those incorporating six spectral bands, yielded the best model performance for most of the soil properties. The inclusion of SWIR bands generally contributed to an improvement in prediction accuracy. On the other hand, PlanetScope, despite having a higher spatial resolution of 3 meters, did not outperform Sentinel-2 MSI or Landsat-8 due to its limited spectral range, particularly the absence of SWIR bands. The soil property maps generated from these satellite images varied in spatial detail and accuracy, with significant implications for soil classification and management. For example, the delineation of soil classes such as Nitisol based on clay content and the identification of variations in organic matter and iron content were influenced by the choice of satellite data. The study concludes that satellite images, especially multi-temporal composites, are effective tools for digital soil mapping, providing valuable information for soil classification and management. However, the utility of these satellite images depends on the specific application and the required spatial resolution. Integration of field data remains crucial for capturing critical subsurface variability that satellite-derived data may not fully represent, ensuring accurate soil property mapping and classification.

Pham et al. (2021)addresses critical gaps in soil property analysis through advanced interactive visualization techniques and machine learning approaches, informed by extensive stakeholder engagement including interviews with 102 professionals in the proximal sensor field. The research identified five key objectives: developing typical visualizations for chemical measurement data, creating intelligent visual recommendation systems, implementing real-time error detection for proximal sensors, advancing machine learning components for device calibrations, and establishing predictive models for soil property estimation from spectral data.

The visualization framework addresses portable X-ray fluorescence (pXRF) soil profile data analysis through sophisticated interactive solutions including force-directed correlation graphs, scatter plots with linear regression analysis, interpolated contour maps using spherical Kriging algorithms, statistical box plots for horizon-based distribution analysis, and innovative 3D visualization systems inspired by medical imaging techniques. Five critical analysis tasks were systematically addressed: providing comprehensive elemental overviews, quantifying correlations between chemical elements, comparing spatial distributions, analyzing statistical distributions across pedological horizons, and detecting outlying data points caused by field scanning errors.

Simultaneously, the study explores machine learning and deep learning approaches for predicting soil properties from visible and near-infrared (Vis-NIR) spectral data using the extensive ICRAF-ISRIC global soil spectral library containing 4,437 samples from 785 soil profiles across 58 countries. The spectral library encompasses wavelengths from 350-2500 nm across 216 wavebands, focusing on predicting pH_H_2O and pH_KCl values as alternatives to time-consuming laboratory procedures. Data preprocessing employed Savitzky-Golay transformation to enhance signal-to-noise ratios and capture reflectance changes between consecutive wavebands.

Comprehensive model comparison included Partial Least Squares Regression, Random Forest with optimized hyperparameters, Multi-Layer Perceptron with five fully connected layers, and various Convolutional Neural Network architectures including DenseNet and VGG blocks. The novel Residual Dilated Neural Network (RDNet) architecture was developed incorporating WaveNet-inspired dilated convolutions with ResNet skip connections, enabling efficient feature extraction from long spectral sequences while minimizing parameter requirements. RDNet achieved state-of-the-art performance with mean squared error of 0.28 and 0.25, coefficient of determination (R^2) of 0.86 for both pH measurements, and residual prediction deviation of 2.76 and 2.93 for pH_H_2O and pH_KCl respectively, substantially outperforming conventional methods. This integrated approach provides both immediate analytical capabilities for field scientists and robust predictive modeling alternatives to expensive, time-consuming laboratory procedures.

The study Zolfaghari Nia et al. (2022) explored the spatial variability of soil properties in riparian forests and adjacent agricultural lands using ML models. A total of 103 soil samples were collected from the Maroon riparian forest in Iran using the Latin hypercube sampling method Shields and Zhang (2016). Various soil properties such as nitrogen, potassium, organic carbon, C:N ratio, pH, calcium carbonate, sand, silt, clay, and bulk density were analyzed.

To model and map these properties, five ML algorithms were evaluated: Random Forest (RF), Cubist regression tree (RTC), Artificial Neural Networks (ANN), k-Nearest Neighbors (KNN), and Classification and Regression Trees (CART). The study used remote sensing data (MODIS, Landsat-8, Sentinel-2), digital elevation models (DEM), and climate variables as ancillary data. The Boruta algorithm Kursa and Rudnicki (2010) was applied to identify the most significant predictors.Key findings include:

- RF performed best for predicting pH, nitrogen, potassium, and bulk density.
- RTC outperformed others for organic carbon, C:N ratio, phosphorous, and clay content.
- ANN was most accurate for calcium carbonate, sand, and silt.

The results highlight the importance of selecting model-specific approaches for different soil properties and demonstrate the effectiveness of ML combined with geospatial data for digital soil mapping (DSM). The study emphasizes that DSM can serve as a cost-effective and accurate tool for environmental planning and monitoring in riparian ecosystems.

Peng et al. (2022) presents a comprehensive methodology for quantitative soil fertility assessment utilizing advanced crop spectral variables derived from high-resolution Sentinel-2 satellite imagery, representing a paradigm shift from traditional laboratory-based soil analysis to remote sensing applications. Conducted in the agriculturally significant Conghua District of Guangzhou, Guangdong Province, China (113°17 E–114°04 E, 23°22 N–23°56 N), researchers systematically collected 150 strategically distributed soil samples using stratified random sampling methodology considering diverse land units, soil types, land-use patterns, and agricultural facility construction levels across the 205 km^2 of arable land.

The comprehensive soil analysis protocol examined five critical soil properties: pH levels (ranging 4.90-8.20), soil organic matter content (6.42-68.90 g/kg), total nitrogen concentrations (0.37-2.14 g/kg), available phosphorus levels (6.80-140.8 mg/kg), and available potassium content (2.00-235.00 mg/kg). These measurements were integrated using fuzzy mathematical approaches following DB43/T 2087-2021 regulations to calculate a comprehensive Soil Fertility Index (SFI) incorporating weight coefficients and membership degrees for each indicator. The innovative approach combines Extreme Gradient Boosting (XGBoost) algorithm for optimal variable selection with Backpropagation Neural Network (BPNN) for sophisticated fertility estimation modeling.

From an initial set of 27 crop spectral variables calculated using the Google Earth Engine platform, the XGBoost algorithm with optimized parameters The model was trained using a learning rate ($\eta = 0.4$), which allows it to update weights relatively quickly during the training process, enabling faster convergence but also increasing the risk of overshooting the optimal solution. To capture complex patterns in the data, the maximum depth of each tree was set to 10 ($\mathsf{max\_depth} = 10$), providing the model with enough capacity to learn detailed interactions between features. Additionally, the training process was carried out over 150 boosting rounds ($n_{\text{round}} = 150$), allowing the model to incrementally improve its performance with each iteration. Identified six preliminary variables, subsequently refined using Variance Inflation Factor ($\mathsf{VIF} < 10$) analysis to eliminate multicollinearity. Five optimal crop spectral variables were ultimately selected: inverted red-edge chlorophyll index (IRECI), chlorophyll vegetation index (CVI), normalized green–red difference index (NGRDI), red-edge position (REP), and triangular greenness index (TGI). This represents the pioneering application of red-edge spectral indices for soil fertility evaluation, leveraging Sentinel-2's unique red-edge bands (705–783 nm) that provide enhanced vegetation stress detection capabilities.

The BPNN model architecture featured 11 hidden layer neurons, 5000 training iterations, and 0.01 learning rate, achieving superior performance with coefficient of determination (R^2) Onyutha (2020) of 0.66, root mean square error (RMSE) Hodson (2022) of 0.17, concordance correlation coefficient (CCC) King et al. (2007) of 0.81, and ratio of performance to interquartile range (RPIQ) Breure et al. (2022) of 1.16 on validation data, substantially outperforming multiple linear regression approaches (R^2=0.02, RMSE=0.28). Regional-scale validation using six Sentinel-2 images spanning September-November 2017 rice growing seasons demonstrated robust model transferability with R^2 of 0.62 and RMSE of 0.09, confirming practical applicability for precision agriculture implementations. The successfully captured nonlinear relationships between spectral variables and soil fertility provide rapid, cost-effective alternatives to time-consuming laboratory procedures, enabling real-time agricultural decision-making and sustainable farming practices across large geographical areas.

El Behairy et al. (2024b) focuses on the use of machine learning, specifically ANN, to accurately predict soil quality indices in arid regions. The authors utilized extensive soil data, employing a MATLAB-based program to process and analyze the information efficiently. The ANN model demonstrated high predictive accuracy, achieving coefficients of determination of approximately 0.97 for training and 0.98 for testing datasets. The findings indicated that a significant portion of the soil samples (36.93%) fell into the very high-quality category, while other quality categories showed varied distributions, highlighting the importance of soil features like pH, salinity, and calcium content in determining soil quality.

The study indicates that traditional methods of soil assessment are often inefficient and that machine learning approaches can provide more reliable results. By utilizing 306 soil samples from three distinct regions, the researchers established a comprehensive SQI database consisting of chemical, physical, and fertility indicators, which are essential for understanding soil characteristics and management.

The study illustrates the potential of machine learning in agricultural practices, offering a robust methodology for predicting soil quality that can be applied to other regions. The authors suggest that regular soil quality evaluations are crucial for improving crop yields and addressing food security concerns. Future research should explore various algorithms and activation functions in ANN models, as well as incorporate biophysical and socio-economic factors to enhance the understanding of soil quality dynamics. The proposed approach serves as a valuable tool for decision-makers in optimizing soil management practices to meet the challenges of sustainable agriculture.

2.4 Deep Learning Methods

In Padarian et al. (2019), the authors investigate the use of Diffuse Reflectance Infrared Spectroscopy (DRIS) to quickly get soil information, whether in the field or the lab. They point out that the growing global interest in vis-NIR spectroscopy has led to the creation of regional and even global soil spectral libraries. These large datasets can be tough to analyze using traditional methods, but the authors believe that deep learning, especially convolutional neural networks (CNNs), offers a promising solution for processing this data more efficiently.

The study focuses on using CNNs to predict various soil properties directly from raw soil spectra, without needing to pre-process the data. They tested this approach on the LUCAS soil database, which contains around 20,000 soil observations from Europe, covering a wide range of physicochemical and biological properties. The authors represented the soil spectral data as 2D spectrograms, which show reflectance values as a function of wavelength and frequency. This technique allowed them to train the CNN in a multi-task setting, as shows Figure 2.1,Common layers represent the layers shared by all the predicted properties. Each branch, one per predicted soil property, correspond to a series of one convolutional layer (BN: bottle-neck layer, which reduces the dimensionality of the data) and a fully-connected layer of size 1, which corresponds to the final prediction. The the model was able to predict six soil properties simultaneously: organic carbon (OC), cation exchange capacity (CEC), clay content, sand content, pH, and total nitrogen (N).

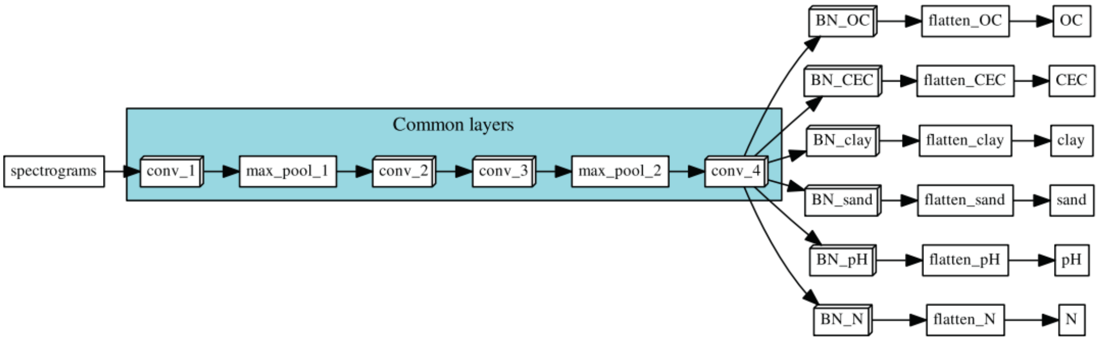


Figure 2.1: Architecture of the multi-task network.

The results were impressive. The CNN outperformed traditional methods like PLS regression and Cubist regression trees, especially in multi-task learning. For example, in predicting soil organic carbon, the multi-task CNN reduced prediction errors by 87% compared to PLS and 62% compared to Cubist. The study shows that CNNs are highly effective for modeling soil spectral data, especially when trained on large datasets. Given their high accuracy, CNNs are seen as an ideal tool for analyzing complex soil data.

In Sumathi et al. (2023), the study proposes the Improved Soil Quality Prediction Model using Deep Learning (ISQP-DL) to enhance soil quality prediction for smart agriculture in Coimbatore and Erode districts of Tamil Nadu, India. ISQP-DL model consists of Deep Neural Network Regression (DNNR) architecture with some hidden units layers to investigate chemical (e.g., organic carbon, phosphorus, potassium), physical (e.g., pH), and biological (e.g., natural manure) properties of the soil.

Soil laboratory values (2016–2020) are pre-processed, trained, and tested using classification grouped into six fertility classes of Very-Less (A), Less (B), Medium (C), Modest (D), High (E), and Max-rate (F). The soil features are processed by the model that classifies soil quality at 96.7% accuracy, better than traditional models such as ANN and KNN. It also utilizes rectified linear activation function and Levenberg-Marquardt optimization Hemmati-Sarapardeh et al. (2020) for lesser training time and better generalization.

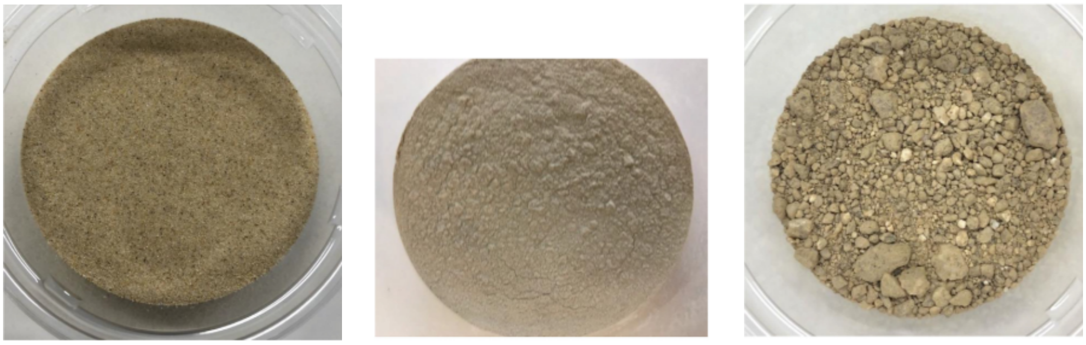
The ISQP-DL model not only reduces computational complexity but also produces SQ Reports supporting data-driven crop planning and fertilizer recommendations. The system further incorporates IoT and cloud infrastructure for real-time monitoring of agriculture. Integration with automated irrigation in the future is suggested by the authors to further enhance agricultural decision-making.

Folorunso et al. (2023) carried out a systematic review on the application of machine learning models in predicting different nutrient attributes of soil, which is an important area in precision agriculture. The authors have focused on the huge strides achieved by Digital Soil Mapping (DSM), which, in their opinion, were enabled by the advanced functionalities of different machine learning methodologies, including Random Forest, Support Vector Machines, and Deep Neural Networks, to improve soil quality assessments, crop yield prediction, and resource management practices. The study emphasized the fact that analysis of traditional soil data poses computational challenges, especially when large datasets are involved, as in the case of recently developed global and regional soil spectra libraries.

Key in the review was the innovative use of Diffuse Reflectance Infrared Spectroscopy—commonly known as DRIS—for fast gathering of field or lab soil information by using deep learning in analyzing raw soil spectral data; authors recommended using CNN. They have done this by applying the above approach on the LUCAS soil database, which includes 20,000 soil observations across Europe. They represented the spectral data of the soil into 2D spectrograms and then trained CNNs in a multitask framework for the prediction of the following six soil properties: organic carbon, cation exchange capacity, clay content, sand content, pH, and total nitrogen.

The review also noted some of the limitations, including data availability, technological, and infrastructural barriers that slow the uptake of these technologies, especially in developing countries. The authors finally provide a framework for future research, which emphasizes the need for state-of-the-art soil information systems, seamless data integration, and tailored machine learning approaches to improve agricultural productivity worldwide.

Inazumi et al. (2020) use deep learning to automate soil classification, the data were collected by images of soil were taken using a smartphone camera and resized to 56x56 pixels for processing, as presented in Figure 2.2.



(a) Image of sand (b) Image of clay (c) Image of gravel

Figure 2.2: Examples of soil image used to develop AI program for soil classification. Inazumi et al. (2020)

This study employs a CNN to classify three soil types: clay, sand, and gravel. The model, as shown by Figure 2.3, was trained using 1,000 images (400 clay, 400 sand, 200 gravel).

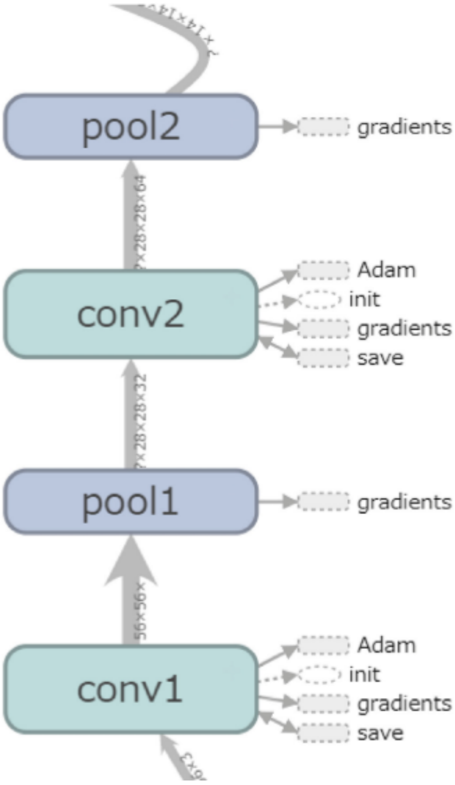


Figure 2.3: Image of model's multi-layer structure

The model achieved an accuracy of 86% on training data and 77% on verification data.Sand was identified with high recall (1.0), but gravel had lower recall (0.54), often being mistaken for sand or clay.

Table 2.1: Recap. of the literature review on Soil Quality Prediction (sorted by year)

Reference	**Methods**	**Dataset**	**Metrics & Results**
Padarian et al. (2019)	CNN on raw DRIS spectra as 2D spectrograms	LUCAS (20,000 soil samples)	87% error reduction (vs. PLS); predicted 6 properties at once
Du et al. (2020)	Micro-CT, segmentation, AVIZO 3D rendering	Zhongcha 108 tea stalks (3 depths × 3 angles)	35° angle optimal; MC-SARCN up to 22.83%, VRCN to 5.76%
Inazumi et al. (2020)	CNN trained on 56×56 images (clay, sand, gravel)	1000 images (400 clay, 400 sand, 200 gravel)	Accuracy: 86% (train), 77% (test); recall 1.0 for sand, 0.54 gravel
Pham et al. (2021)	PLSR, RF, MLP, CNN, Residual Dilated CNN (RDNet)	ICRAF-ISRIC (4,437 samples, 58 countries)	RDNet: $R^2 = 0.86$, MSE = 0.25–0.28, RPD 2.8–2.9
Silvero et al. (2021)	Cubist regression; SYSI and vis-NIR+SWIR bands	120 samples, Brazil + PlanetScope, Sentinel-2, Landsat-8	Sentinel-2 MSI (6 bands) gave best performance; SWIR critical
Peng et al. (2022)	XGBoost for variable selection, BPNN, fuzzy logic	150 samples (Guangzhou, China) + Sentinel-2	$R^2 = 0.66$, RMSE = 0.17, CCC = 0.81; regional $R^2 = 0.62$
Zolfaghari Nia et al. (2022)	RF, RTC, ANN, KNN, CART; Boruta feature selection	103 soil samples (Iran) + MODIS, Sentinel-2, DEMs	RF best for pH/N/K; RTC for OC/clay; ANN for CaCO , sand, silt
Folorunso et al. (2023)	RF, SVM, DNN, CNN with DRIS; multitask framework on 2D spectrograms	LUCAS soil database (20,000 observations, Europe)	CNN predicted 6 properties: OC, CEC, clay, sand, pH, nitrogen
Sumathi et al. (2023)	DNNR, RLAF, Levenberg–Marquardt, 10-fold cross-validation	Soil samples from Tamil Nadu, India (6 quality levels)	Accuracy: 96.7%; high F1-score, outperforming ANN and KNN
El Behairy et al. (2024b)	Artificial Neural Networks via MATLAB	306 soil samples (3 regions, Egypt)	R^2 0.97 (train), 0.98 (test); 36.9% were very high SQI

2.5 Conclusion

This chapter contained a thorough analysis of the key components and methods involved in soil quality analyses, the surveyed literature show that a clear evolution from traditional laboratory-based methods to sophisticated machine learning and deep learning solutions that use advanced imagine technologies.

In the following chapter, we shift from theoretical exploration to practical implementation. The experimental chapter presents the methodological framework adopted for this study, detailing the datasets used, preprocessing steps, and the design of the machine learning models employed for soil quality prediction and plant suggestion.

CHAPTER 3

EXPERIMENTS

3.1 Introduction

In the previous chapter, we presented, a literature reviewof ML and DL based techniques for Soil Quality Prediction. Based on all these works, we present in this chapter our experiment for SQP system. We start by description of the datasets, going through the pre-processing steps adopted in addition to the architecture of our models, and finally our results. Our work is divided into two experiments in the first, we conduct a comparative study of 4 ML models for SQI prediction and mapping. Based on the results, we select the most effective model and use it to generate the SQI map. In the second experiment, we trained our RF model for SQI prediction, then, in order to optimize the model, we create a model that suggests plants based on SQI, weather and expert rules. Finaly a web application is proposed in order to facilitate the usage of our models.

3.2 Experiment I: Mapping SQI using best regressor

In order to create a map of soil quality, we first study four ML/DL regression techniques[1], then choose the best among them. The process is shown in Figure 3.1. In the sequel, we describe each stage.

3.2.1 Dataset collection

We experiment with the SoilGrids dataset provided by ISRIC-World Soil Information Poggio et al. (2021). It offers consistent quantitative information about soil characteristics across the globe, making it suitable for ML applications in soil quality and construction suitability prediction.

[1] This part is presented as conf. paper on *AISTC'2025* : International Conference on Artificial Intelligence, Smart Technologies and Communications on April 14-15, 2025 at Chl

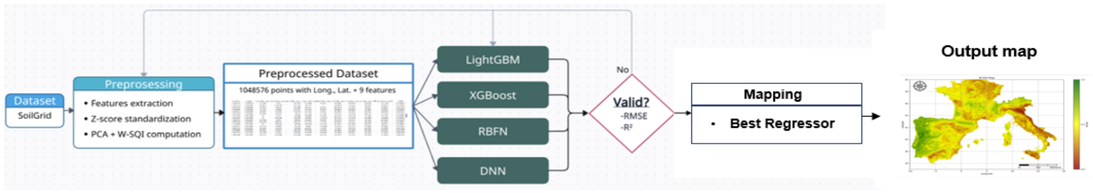


Figure 3.1: The performed SQP pipeline

3.2.2 Preprocessing

From SoilGrids dataset a tabular dataset is created to facilitate analysis. First, each cell of the grid is represented by its center as a point with (longitude, latitude) and nine important soil characteristics to compute SQI, as described in Table 3.1.

Table 3.1: Soil features description.

Feature	Unit	Description
Bulk density	(cg/cm^3)	Mass of soil per volume unit, indicating compaction and porosity.
Cation exchange capacity at pH 7	(mmol(c)/kg)	Soil's ability to retain and exchange nutrients.
Coarse fragments	(cm^3/dm^3)	Volume of large soil particles (gravel, stones) affecting drainage.
Clay content	(g/kg)	Fine soil particles that retain water but reduce aeration.
Nitrogen	(cg/kg)	Essential nutrient for plant growth and chlorophyll production.
pH water	(pH $\times 10$)	Soil acidity or alkalinity, affecting nutrient availability.
Sand	(g/kg)	Large soil particles improving drainage but reducing water retention.
Silt	(g/kg)	Medium-sized particles balancing moisture retention and drainage.
Soil organic carbon	(dg/kg)	Carbon in organic matter, vital for soil fertility.

Due to differences in units, standardization is performed using z-score normalization. To compute an accurate SQI, the improved method proposed by Damiba et al. (2024) is considered, where a Weighted SQI is calculated, based on PCA to extract the most significant indicators with eigenvalues greater than 1 were considered to account for a significant amount of variance. At the end of this stage, we get a dataset with 2,316,650 points, 11 features and the W-SQI target.

3.2.3 Modeling

The Weighted SQI and predictor variables are used to train and test four ML models, namely LightGBM, XGBoost, Radial Basis Function Network (RBFN), and Deep Neural Network (DNN). To provide a clear understanding of the model design, the architectural overview of each model is illustrated in Figures 3.2 3.3 3.4 3.5. These diagrams highlight the structural flow and key components involved in the implementation of each approach.

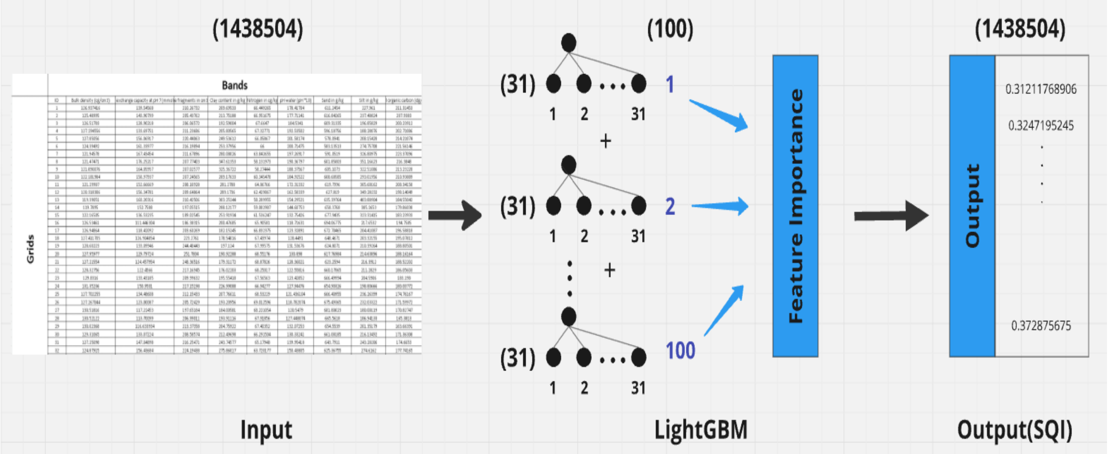


Figure 3.2: Architecture of the LightGBM model

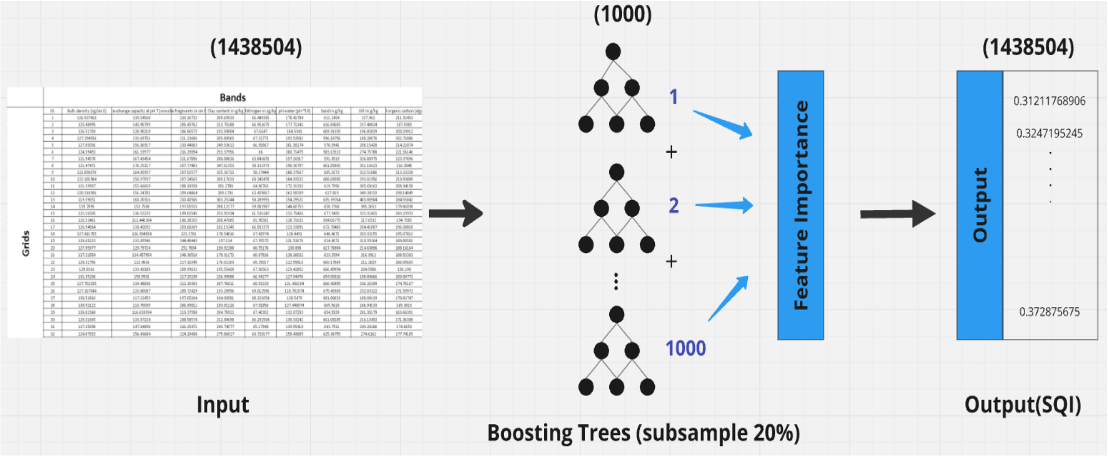


Figure 3.3: Architecture of the XGBoost model

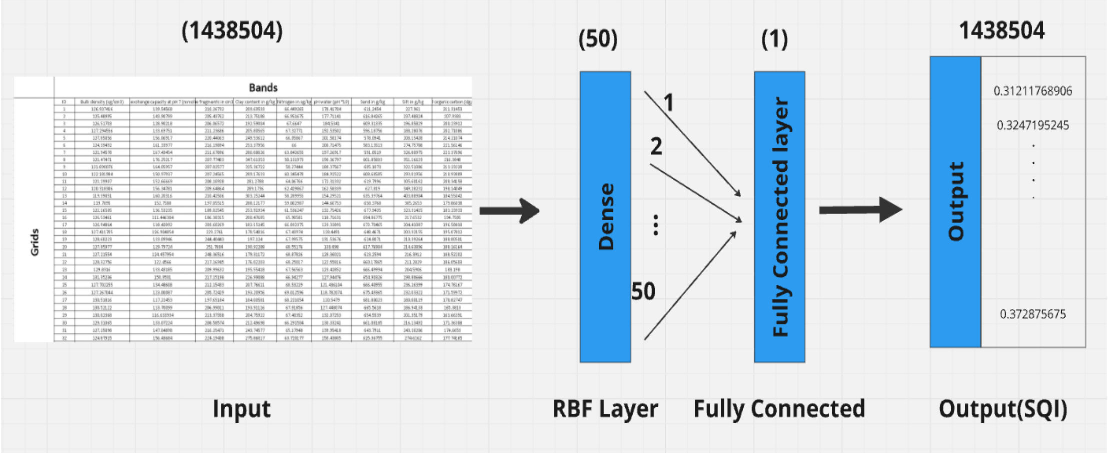


Figure 3.4: Architecture of the Radial Basis Function Network (RBFN)

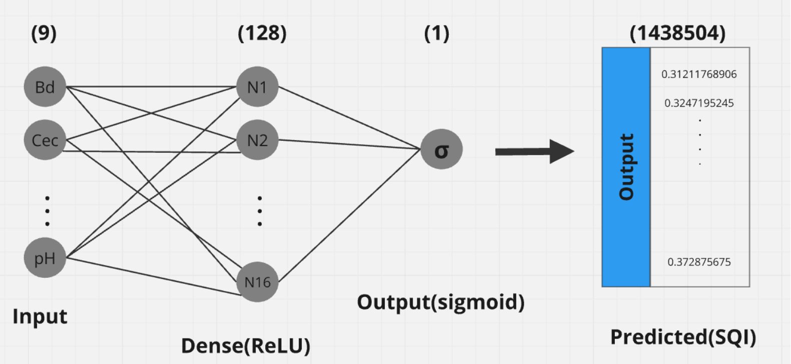


Figure 3.5: Architecture of the Deep Neural Network (DNN)

We selected these algorithms because of their complementary strengths in regression tasks. RBFN is suitable for small datasets and picking up local patterns. LightGBM and XGBoost are gradient boosters which specialize in speed, stability, and performance when dealing with structured data. DNN was employed to evaluate the strength of deep learning in capturing complex, nonlinear relationship among soil parameters.

Hyperparameters for each model were chosen after a few trial runs with varying configurations to get the best performance. Regularization techniques such as dropout (DNN), subsampling (LightGBM), and `colsample_bytree` (XGBoost) were applied to prevent overfitting. Learning rates (0.05 for LightGBM/XGBoost, 0.001 for DNN) were such that stable convergence could be obtained. Number of epochs and model-specific parameters (e.g., number of leaves in LightGBM, centers in RBFN) were tried and tuned for balancing accuracy, efficiency, and generalization.

3.2.4 Results and Discussion

The result presented in Table 3.2 indicate that XGBoost achieved the highest R^2 score of 0.9768, making it the most effective model to predict soil quality in this setting. LightGBM achieved an R^2 score of 0.9043, demonstrating its strong predictive capability while maintaining computational efficiency with a significantly lower execution time of 1.78 seconds. DNN, despite achieving a comparable R^2 score of 0.8981, required a much longer execution time of 1176.19 seconds, making it less efficient for this task. Ridge regression with an RBF kernel, on the other hand, had the lowest performance with an R^2 score of 0.7925, indicating weaker predictive accuracy compared to the other models.

Table 3.2: Comparison of the four models based on execution time, RMSE, and R^2 score.

Model	Time (s)	RMSE	R^2
RBFN	147.10	0.03	0.79
DNN	1176.19	0.02	0.90
LightGBM	1.78	0.02	0.90
XGBoost	3.54	0.02	**0.98**

Using geospatial techniques and ML models, it is possible to generate high-resolution soil quality maps that highlight areas of concern and potential for improvement. Figure 3.6 illustrates the spatial distribution of the Weighted SQI across the Mediterranean and surrounding regions, as predicted by the XGBoost model.

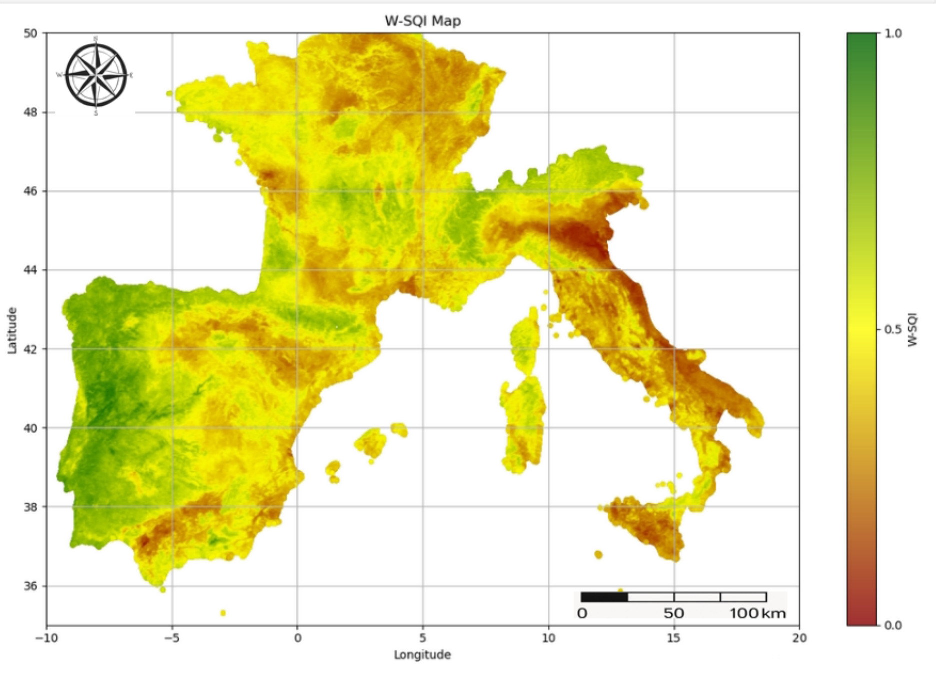


Figure 3.6: Soil Quality Index Map.

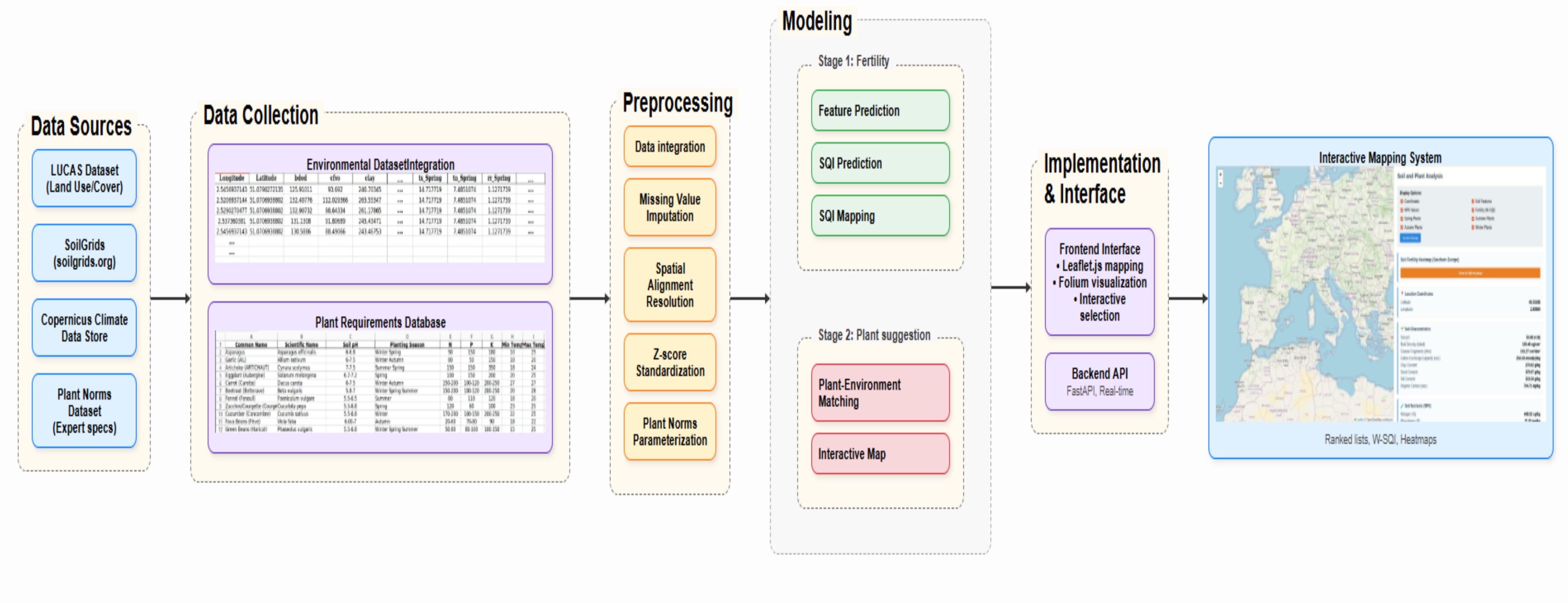
The map employs a continuous color gradient ranging from deep red to green, where green shades signify areas with higher soil quality (W-SQI values up to 0.55), and red hues indicate regions with comparatively lower soil quality (W-SQI values closer to 0.30). Intermediate colors, such as yellow and orange, represent moderate soil quality levels. From the visualization, it is evident that regions in southwestern Europe, particularly parts of Spain and southern France, display higher W-SQI values, indicated by the prevalence of green shades. In contrast, areas in central and southern Italy, as well as coastal zones along the Adriatic Sea, show lower soil quality, represented by red and orange tones. This spatial heterogeneity in soil quality reflects the influence of various geographical and environmental factors, such as topography, climate, land use, and anthropogenic pressures. Notably, mountainous and coastal areas appear more prone to lower soil quality, possibly due to erosion, land degradation, or intensive agricultural activities. Such a map is invaluable for stakeholders in precision agriculture, land conservation, and sustainable land management, providing critical information to optimize resource allocation, improve crop productivity, and implement targeted conservation strategies. Furthermore, identifying areas with declining soil quality can guide policymakers in prioritizing soil restoration and environmental protection efforts.

3.3 Experiment II: extending SQI Mapping with Plants Suggestion

Implementation setup Implementations are carried out using the Keras API and TensorFlow backend. Experiment is executed on a system with Intel i7 CPU, 32GB RAM, and NVIDIA RTX 4060 GPU under Ubuntu 24.04.1 LTS.

To extend our work beyond soil quality evaluation, this second experiment aims to identify and map the most suitable plant species for a given region using machine learning techniques.

The methodological pipeline is summarized in Figure 3.7, encompassing dataset preparation, preprocessing, feature selection, model training, and spatial prediction.


Data Sources
LUCAS Dataset (Land Use/Cover)
SoilGrids (soilgrids.org)
Copernicus Climate Data Store
Plant Norms Dataset (Expert specs)
Data Collection
Environmental DatasetIntegration
Plant Requirements Database
Preprocessing
Data integration
Missing Value Imputation
Spatial Alignment Resolution
Z-score Standardization
Plant Norms Parameterization
Modeling
Stage 1: Fertility
Feature Prediction
SQI Prediction
SQI Mapping
Stage 2: Plant suggestion
Plant-Environment Matching
Interactive Map
Implementation & Interface
Frontend Interface
• Leaflet.js mapping
• Folium visualization
• Interactive selection
Backend API
FastAPI, Real-time
Interactive Mapping System
Soil and Plant Analysis
Ranked lists, W-SQI, Heatmaps

Longitude	Latitude	bdod	cfvo	clay	...	tx_Spring	tn_Spring	rr_Spring	...
2.5456937143	51.0790272135	125.91011	93.692	240.70345	...	14.717719	7.4851074	1.1271739	...
2.5208937144	51.0706938802	132.40776	112.020366	265.53347	...	14.717719	7.4851074	1.1271739	...
2.5290270477	51.0706938802	132.90732	98.64334	261.17065	...	14.717719	7.4851074	1.1271739	...
2.537380381	51.0706938802	131.1308	91.80689	245.43471	...	14.717719	7.4851074	1.1271739	...
2.5456937143	51.0706938802	130.5886	88.49066	243.46753	...	14.717719	7.4851074	1.1271739	...
...									
...									

Common Name	Scientific Name	Soil pH	Planting Season	N	P	K	Min Temp	Max Temp
Asparagus	Asparagus officinalis	6-6.8	Winter Spring	90	150	180	10	25
Garlic (AIL)	Allium sativum	6-7.5	Winter Autumn	80	50	150	10	20
Artichoke (ARTICHAUT)	Cynara scolymus	7-7.5	Summer Spring	150	150	350	18	24
Eggplant (Aubergine)	Solanum melongena	6.7-7.3	Spring	100	150	200	20	25
Carrot (Carotte)	Dacus carota	6-7.5	Winter Autumn	150-200	100-120	200-250	27	27
Beetroot (Betterave)	Beta vulgaris	5.8-7	Winter Spring Summer	150-200	100-120	200-250	20	28
Fennel (Fenouil)	Foeniculum vulgare	5.5-6.5	Summer	80	110	120	18	20
Zucchini/Courgette (Courge	Cucurbita pepo	5.5-6.8	Spring	120	60	100	23	25
Cucumber (Concombre)	Cucumis sativus	5.5-6.8	Winter	170-200	100-150	200-250	22	25
Fava Beans (Fève)	Vicia faba	6.00-7	Autumn	20-60	70-80	90	18	22
Green Beans (Haricot)	Phaseolus vulgaris	5.5-6.8	Winter Spring Summer	50-80	80-100	100-150	15	25

Figure 3.7: Pipeline for identifying and mapping optimal plant species

3.3.1 Dataset collection

3.3.1.1 Environmental Data

The environmental dataset served as the foundation for spatial analysis, containing comprehensive environmental and climatic information for various geolocations. These spatial variables, along with the soil and environmental parameters, were derived from a combination of the LUCAS (Land Use/Cover Area frame Survey)[2], SoilGrids[3], and climate variables obtained via the Copernicus Climate Data Store[4].

3.3.1.2 Plant Norm Data

The plant norm dataset contained expert-derived specifications for optimal growing conditions of various plant species. This dataset was compiled through consultation with soil and agriculture specialists and illustrated by table 3.3.

Category	**Description**
Species Identification	Common Name: Vernacular plant names Scientific Name: Taxonomic nomenclature following standard botanical classification
Growing Requirements	Soil pH: Optimal pH range for plant growth Planting Season: Recommended planting periods Nutrient Requirements: Optimal levels for Nitrogen (N), Phosphorus (P), and Potassium (K) Temperature Thresholds: Minimum and maximum temperature tolerance limits

Table 3.3: Plant species and their growing requirements

3.3.2 Preprocssing

The preprocessing phase was a crucial step in preparing the data for subsequent modeling tasks. It began with the integration of three primary data sources that are mentioned before. These datasets originated from different platforms and were collected using varying spatial resolutions, coordinate systems, and data formats.This required significant preprocessing to align and standardize the geolocation data before use four steps are preformed:

1. Data Integration and Harmonization: Combining these sources posed several challenges. Most notably, discrepancies in geographic coordinates (longitude and latitude) were encountered due to differences in spatial reference systems and granularity. To ensure spatial compatibility, we employed a nearest neighbor matching strategy. This involved assigning to each soil observation the values of the closest corresponding grid cell from other datasets. This method preserved

[2] `https://esdac.jrc.ec.europa.eu/content/lucas-2009-topsoil-data`

[3] `https://soilgrids.org`

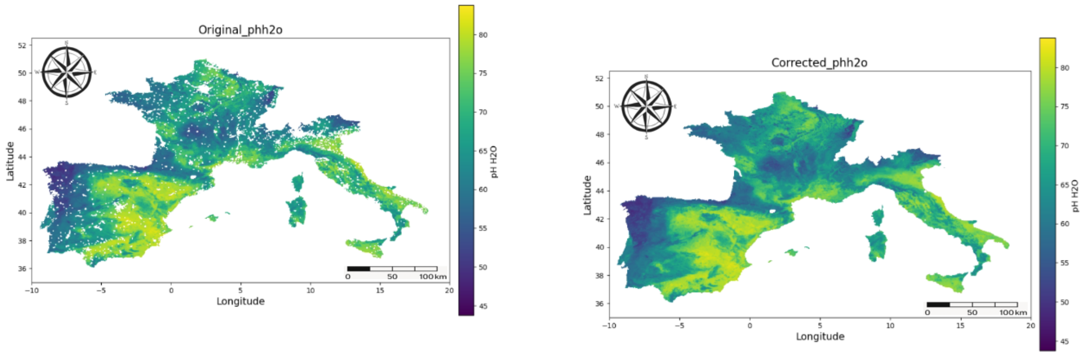
[4] `https://cds.climate.copernicus.eu`

the original resolution of each dataset and avoided the distortions that can arise from interpolation or resampling, thereby ensuring a reliable and coherent spatial integration for subsequent analysis.

2. Spatial Alignment: Even after harmonization, non-coincident spatial points remained due to the nature of data acquisition. To overcome this, nearest neighbor method was used to align the datasets geographically. This step was particularly important to ensure that soil and climate features corresponded accurately to the same physical locations. The resulting aligned dataset allowed for pixel-wise feature extraction, crucial for both prediction and mapping tasks.

3. Missing Value Imputation: All datasets exhibited varying degrees of missing data due to sensor errors, data acquisition gaps, or unreported measurements. Rather than discarding incomplete samples—which would reduce the overall data volume—regression-based imputation techniques were used to estimate missing values.Random Forest Regressor models were trained on correlated features to predict and fill in the gaps.

The implementation of a regression model for missing value imputation can be effectively demonstrated through comparative visualization as indicated by figure 3.8.



(a) Original dataset map with missing pH values

(b) Dataset map after pH imputation

Figure 3.8: Comparison of soil pH maps before and after imputation

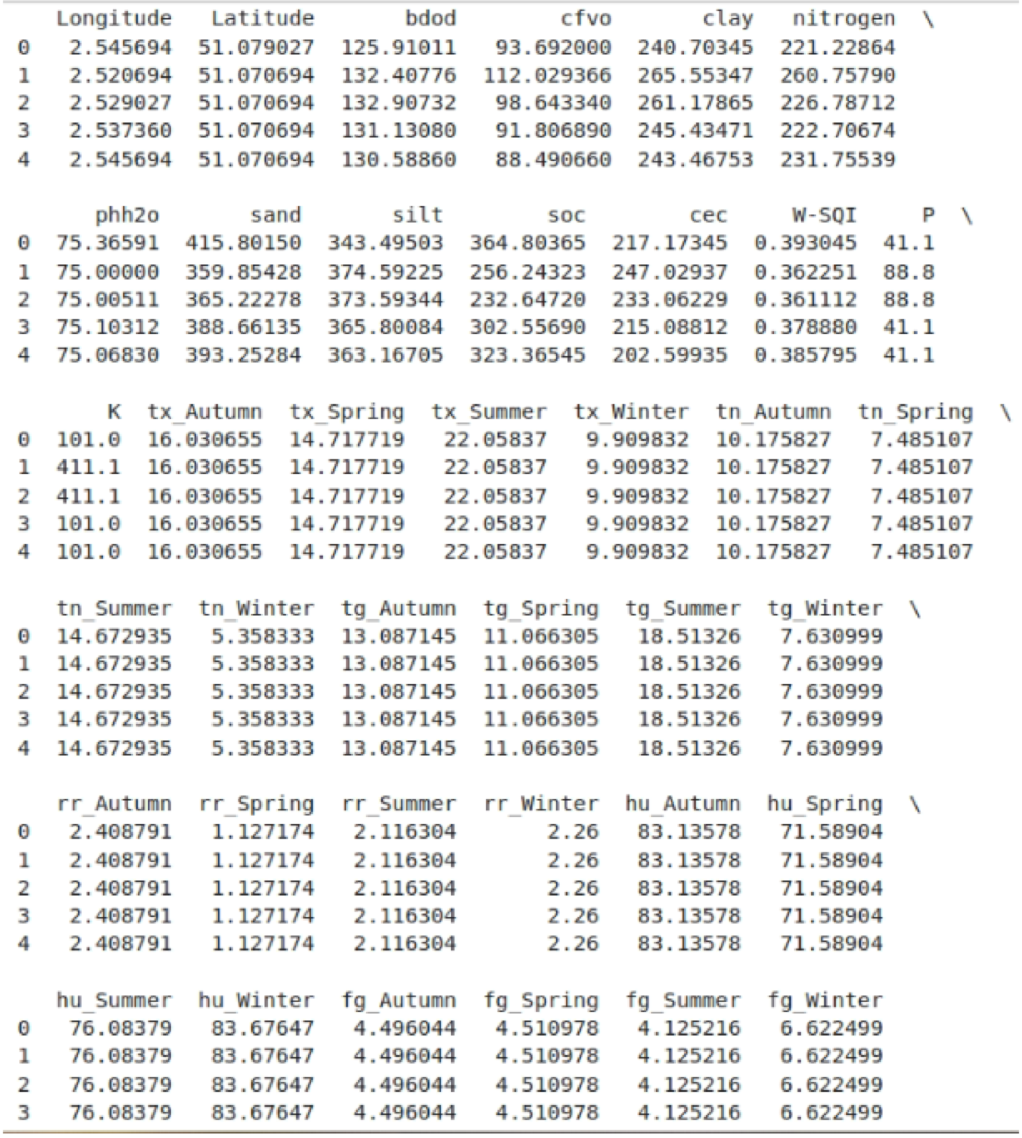
4. Feature Scaling and Normalization: Given the wide disparity in feature units and magnitudes—e.g., pH (unitless), temperature (°C), and nutrient concentrations (mg/kg)—normalization was essential. Z-score standardization was applied to all numerical variables to ensure a mean of zero and unit variance. This scaling ensured that no single feature dominated the learning process.

Together, these preprocessing steps established a consistent, clean, and spatially aligned dataset that served as the foundation for all subsequent analyses. Table 3.4 shows the final dataset after the preprocessing steps.

Table 3.4: Environmental dataset variables

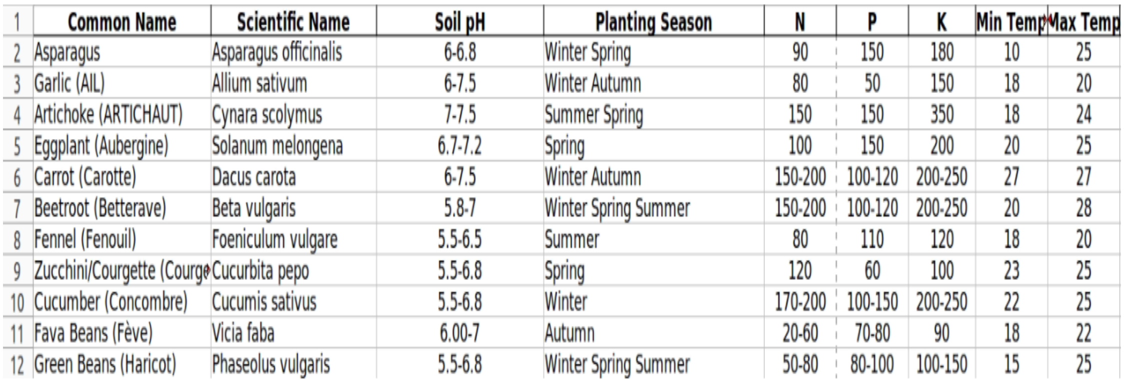
Category	**Description**
Spatial Variables	Geographic coordinates: Longitude and Latitude values defining precise location coordinates
Soil and Environmental Parameters	Phosphorus content (P) Potassium content (K) Bulk density (bdod) Coarse fragments volume (cfvo) Cation exchange capacity (cec) Clay content percentage (clay) Soil pH in water (phh2o) Sand content percentage (sand) Silt content percentage (silt) Soil organic carbon (soc) Nitrogen content (nitrogen) Water Soil Quality Index (W-SQI)
Seasonal Climate Data	Maximum temperature (tx) Minimum temperature (tn) Mean temperature (tg) Precipitation (rr) Humidity (hu) Wind speed (fg)

Table 3.4 provides a sample of the final datasets obtained after the complete preprocessing pipeline.



```
   Longitude   Latitude       bdod        cfvo       clay   nitrogen  \
0   2.545694  51.079027  125.91011   93.692000  240.70345  221.22864
1   2.520694  51.070694  132.40776  112.029366  265.55347  260.75790
2   2.529027  51.070694  132.90732   98.643340  261.17865  226.78712
3   2.537360  51.070694  131.13080   91.806890  245.43471  222.70674
4   2.545694  51.070694  130.58860   88.490660  243.46753  231.75539

      phh2o       sand       silt        soc        cec     W-SQI     P  \
0  75.36591  415.80150  343.49503  364.80365  217.17345  0.393045  41.1
1  75.00000  359.85428  374.59225  256.24323  247.02937  0.362251  88.8
2  75.00511  365.22278  373.59344  232.64720  233.06229  0.361112  88.8
3  75.10312  388.66135  365.80084  302.55690  215.08812  0.378880  41.1
4  75.06830  393.25284  363.16705  323.36545  202.59935  0.385795  41.1

       K  tx_Autumn  tx_Spring  tx_Summer  tx_Winter  tn_Autumn  tn_Spring  \
0  101.0  16.030655  14.717719   22.05837   9.909832  10.175827   7.485107
1  411.1  16.030655  14.717719   22.05837   9.909832  10.175827   7.485107
2  411.1  16.030655  14.717719   22.05837   9.909832  10.175827   7.485107
3  101.0  16.030655  14.717719   22.05837   9.909832  10.175827   7.485107
4  101.0  16.030655  14.717719   22.05837   9.909832  10.175827   7.485107

   tn_Summer  tn_Winter  tg_Autumn  tg_Spring  tg_Summer  tg_Winter  \
0  14.672935   5.358333  13.087145  11.066305   18.51326   7.630999
1  14.672935   5.358333  13.087145  11.066305   18.51326   7.630999
2  14.672935   5.358333  13.087145  11.066305   18.51326   7.630999
3  14.672935   5.358333  13.087145  11.066305   18.51326   7.630999
4  14.672935   5.358333  13.087145  11.066305   18.51326   7.630999

   rr_Autumn  rr_Spring  rr_Summer  rr_Winter  hu_Autumn  hu_Spring  \
0   2.408791   1.127174   2.116304       2.26   83.13578   71.58904
1   2.408791   1.127174   2.116304       2.26   83.13578   71.58904
2   2.408791   1.127174   2.116304       2.26   83.13578   71.58904
3   2.408791   1.127174   2.116304       2.26   83.13578   71.58904
4   2.408791   1.127174   2.116304       2.26   83.13578   71.58904

   hu_Summer  hu_Winter  fg_Autumn  fg_Spring  fg_Summer  fg_Winter
0   76.08379   83.67647   4.496044   4.510978   4.125216   6.622499
1   76.08379   83.67647   4.496044   4.510978   4.125216   6.622499
2   76.08379   83.67647   4.496044   4.510978   4.125216   6.622499
3   76.08379   83.67647   4.496044   4.510978   4.125216   6.622499
```

(a) Final Processed Environmental Dataset



1	Common Name	Scientific Name	Soil pH	Planting Season	N	P	K	Min Temp	Max Temp
2	Asparagus	Asparagus officinalis	6-6.8	Winter Spring	90	150	180	10	25
3	Garlic (AIL)	Allium sativum	6-7.5	Winter Autumn	80	50	150	18	20
4	Artichoke (ARTICHAUT)	Cynara scolymus	7-7.5	Summer Spring	150	150	350	18	24
5	Eggplant (Aubergine)	Solanum melongena	6.7-7.2	Spring	100	150	200	20	25
6	Carrot (Carotte)	Dacus carota	6-7.5	Winter Autumn	150-200	100-120	200-250	27	27
7	Beetroot (Betterave)	Beta vulgaris	5.8-7	Winter Spring Summer	150-200	100-120	200-250	20	28
8	Fennel (Fenouil)	Foeniculum vulgare	5.5-6.5	Summer	80	110	120	18	20
9	Zucchini/Courgette (Courge	Cucurbita pepo	5.5-6.8	Spring	120	60	100	23	25
10	Cucumber (Concombre)	Cucumis sativus	5.5-6.8	Winter	170-200	100-150	200-250	22	25
11	Fava Beans (Fève)	Vicia faba	6.00-7	Autumn	20-60	70-80	90	18	22
12	Green Beans (Haricot)	Phaseolus vulgaris	5.5-6.8	Winter Spring Summer	50-80	80-100	100-150	15	25

(b) Final Plant Norm Dataset

Figure 3.9: Screenshots of the resulting datasets after the complete preprocessing

3.3.3 Modeling

The experimental approach employed a two-stage methodology to achieve a comprehensive plant suitability prediction:

3.3.3.1 Fertility Analysis

The first stage of the system focuses on predicting environmental and soil-related features for unseen geographical coordinates. To achieve this, a separate regression model is trained for each individual feature. Each regressor takes the longitude and latitude as input and outputs a single soil property (e.g., clay content, pH, etc.). The architecture of these regressors is illustrated in the figure 3.10.

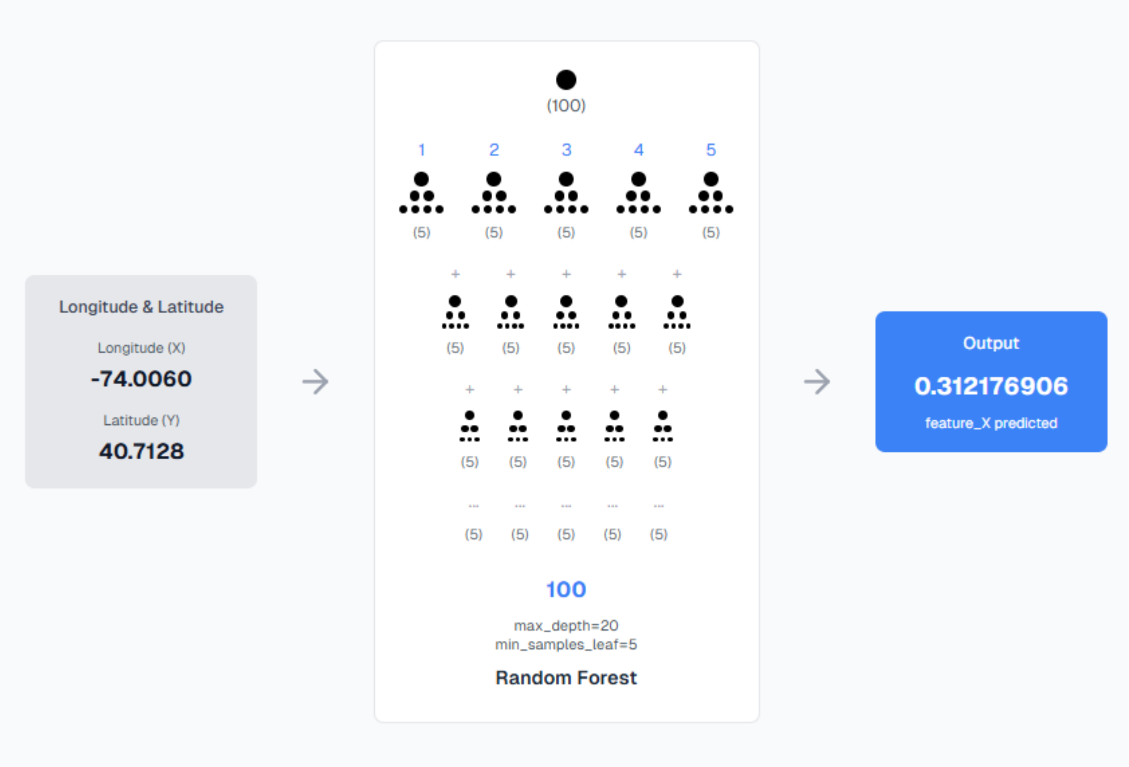


Figure 3.10: Feature Regressor architecture

Once all the target features are predicted, they are collectively passed as input to a second model—a Random Forest Regressor—presented in the figure 3.11. This model is responsible for predicting the Soil Quality Index (SQI) based on the complete set of predicted and geospatial features. The final SQI prediction model is trained on the following input features: `['Longitude', 'Latitude', 'bdod', 'cfvo', 'clay', 'nitrogen', 'phh2o', 'sand', 'silt', 'soc', 'cec', 'P', 'K']`

The Random Forest Regressor is implemented with the following configuration:

```
model = RandomForestRegressor(
    n_estimators=100,
    max_depth=20,
    min_samples_leaf=5,
    n_jobs=-1,
    random_state=42
)
```

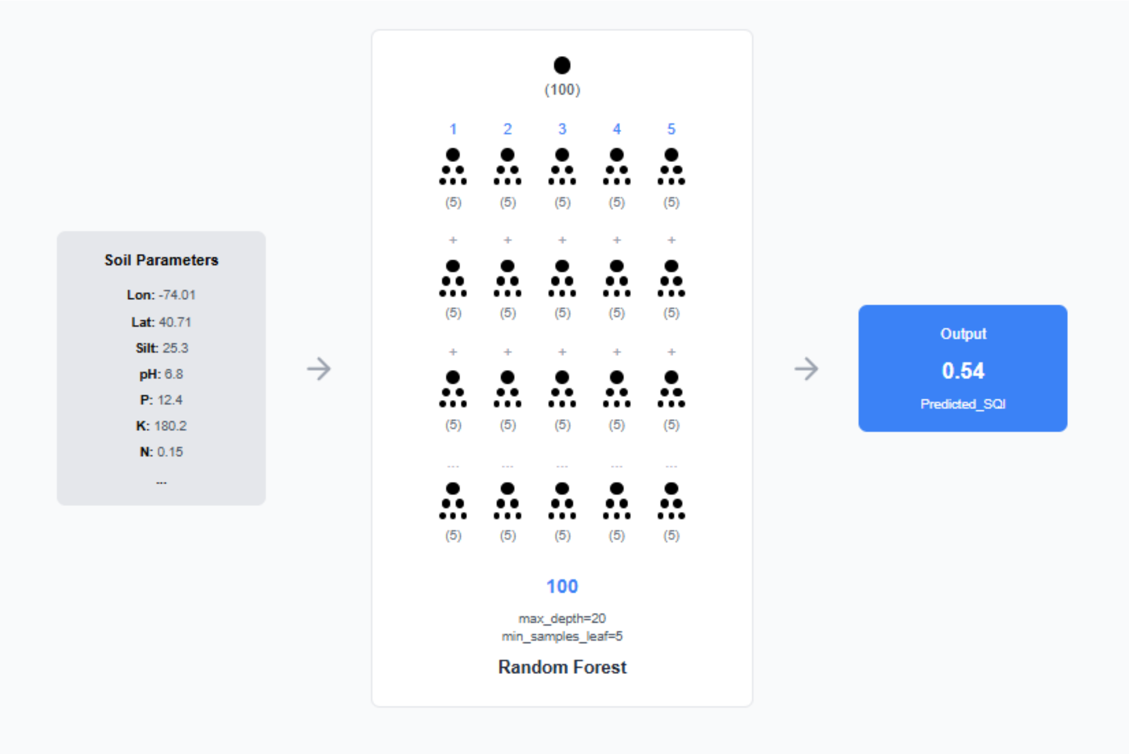
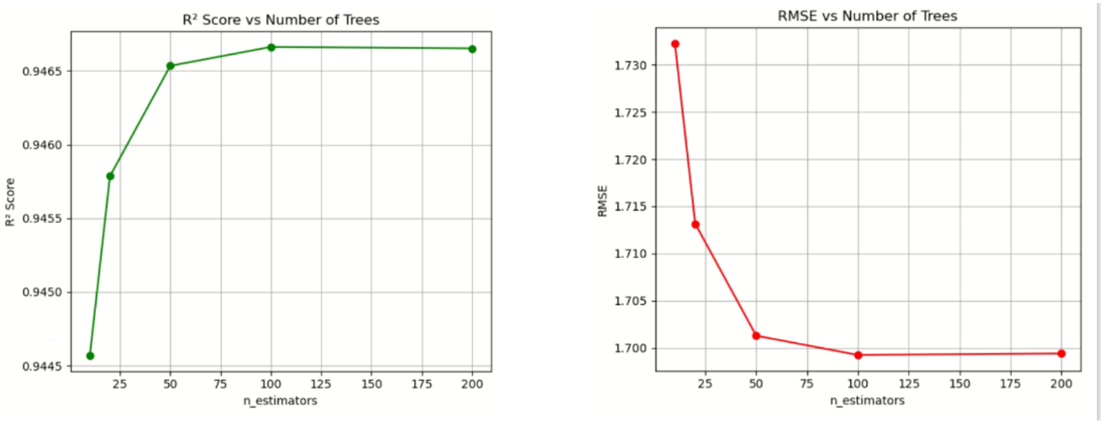


Figure 3.11: SQI Regressor architecture

The Random Forest Regressor exhibited strong predictive capability in estimating environmental parameters across the study area, with R^2 values within acceptable thresholds. Figure 3.12 presents the model's training performance in terms of R^2 and RMSE.



(a) Model R^2

(b) Model RMSE

Figure 3.12: R^2 and RMSE curves

Figure 3.13 illustrate interactive Soil Quality Index (SQI) map created with Folium (an open source library) across the study area, longitude and latitude coordinates locate each prediction. High SQI (green) denotes better soil quality, while low SQI (red) indicates poorer quality.

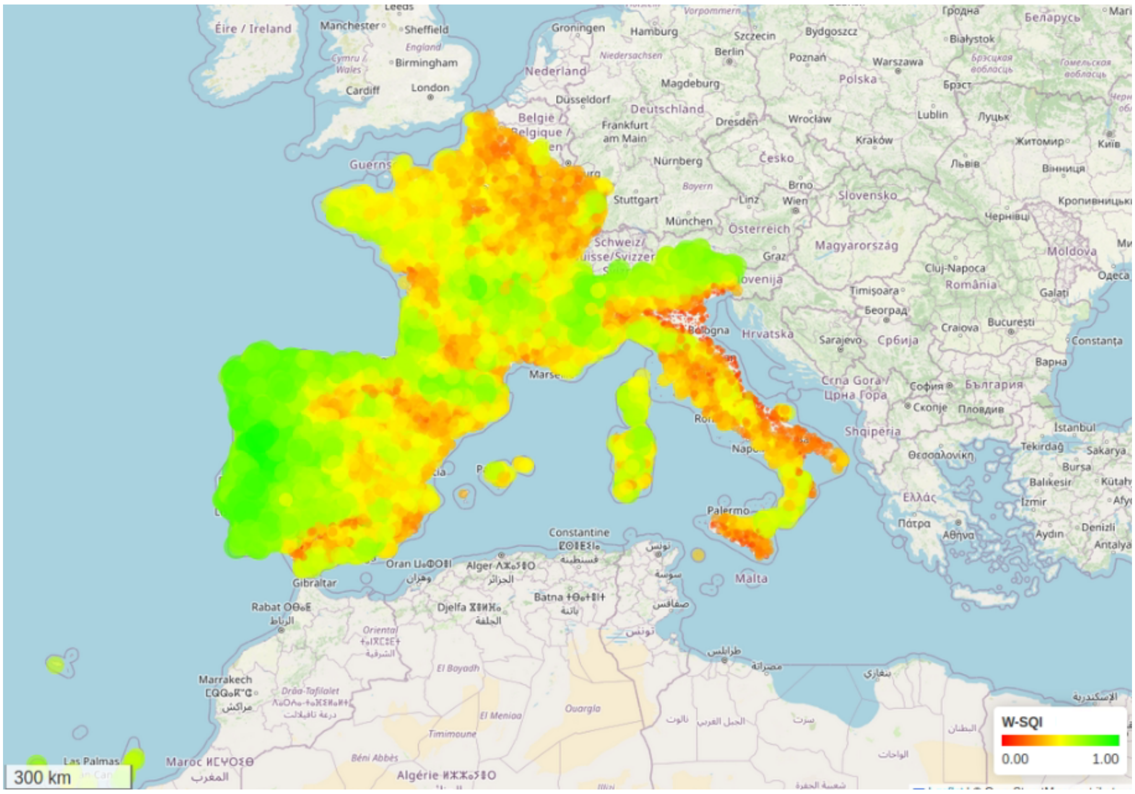


Figure 3.13: Interactive Soil Quality Index (SQI) Map

3.3.3.2 Plant Suggestion

In this part, a plant suggestion is proposed using the **cos similarity**, which is employed to quantify the match between a plant species' optimal growth requirements and the prevailing site and environmental conditions. We represent:

- $\mathbf{s}$: the *site condition vector*, containing parameters such as soil pH, nutrient levels (N, P, K), seasonal minimum and maximum temperatures, etc.
- $\mathbf{p}$: the *plant requirement vector*, encoding the same parameters at their optimal values for a given species.

The cosine similarity between these two vectors is defined as:

$$\cos(\theta) = \frac{\mathbf{s} \cdot \mathbf{p}}{\|\mathbf{s}\| \, \|\mathbf{p}\|} \tag{3.1}$$

Where:

- $s \cdot p$ is the dot product of the two vectors,
- $\|s\|$ and $\|p\|$ are the Euclidean norms (magnitudes) of the vectors,
- θ is the angle between the vectors in a multidimensional space.

A higher cosine similarity value (close to 1) indicates a stronger match between the site conditions and the plant's ideal growing conditions, and thus, a better recommendation.

Because this measure depends solely on the angle between the vectors, it highlights the *relative orientation* of feature patterns (e.g., nutrient ratios) rather than their absolute magnitudes. A value close to 1 indicates a strong correspondence.

The second stage of our pipeline applies this similarity measure to perform plant suggestion based on the previously predicted environmental and soil parameters. Specifically, the system:

1. Retrieves the predicted feature vector $\mathbf{s}$ for the selected location.
2. Compares $\mathbf{s}$ against the plant-norms database $\mathbb{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\}$ where n is the size of the plant database using Equation (3.1).
3. Computes a cosine similarity score for each candidate species.
4. Ranks all species in descending order of similarity.
5. Presents the top three species with the highest scores as the most suitable choices for the given location and season.

Access to additional map features—such as high-resolution soil parameter layers, NPK values, and seasonal vegetation overlays—is determined by the user's subscription level, with higher-tier plans unlocking more detailed data views.

To ensure these recommendations are readily accessible, we constructed an interactive map using the open-source Leaflet library[5] . As shown in Figure 3.14, clicking on any point displays:

- Geographic coordinates,
- Predicted soil characteristics (pH, bulk density, clay, sand, silt, organic carbon, etc.),
- Soil Fertility Index (W-SQI),
- Top three plant suggestions based on cosine similarity.

This interactive map allows stakeholders to explore spatial SQI predictions and receive immediate, tailored plant recommendations, facilitating data-driven agricultural planning.

[5] `https://leafletjs.com/`

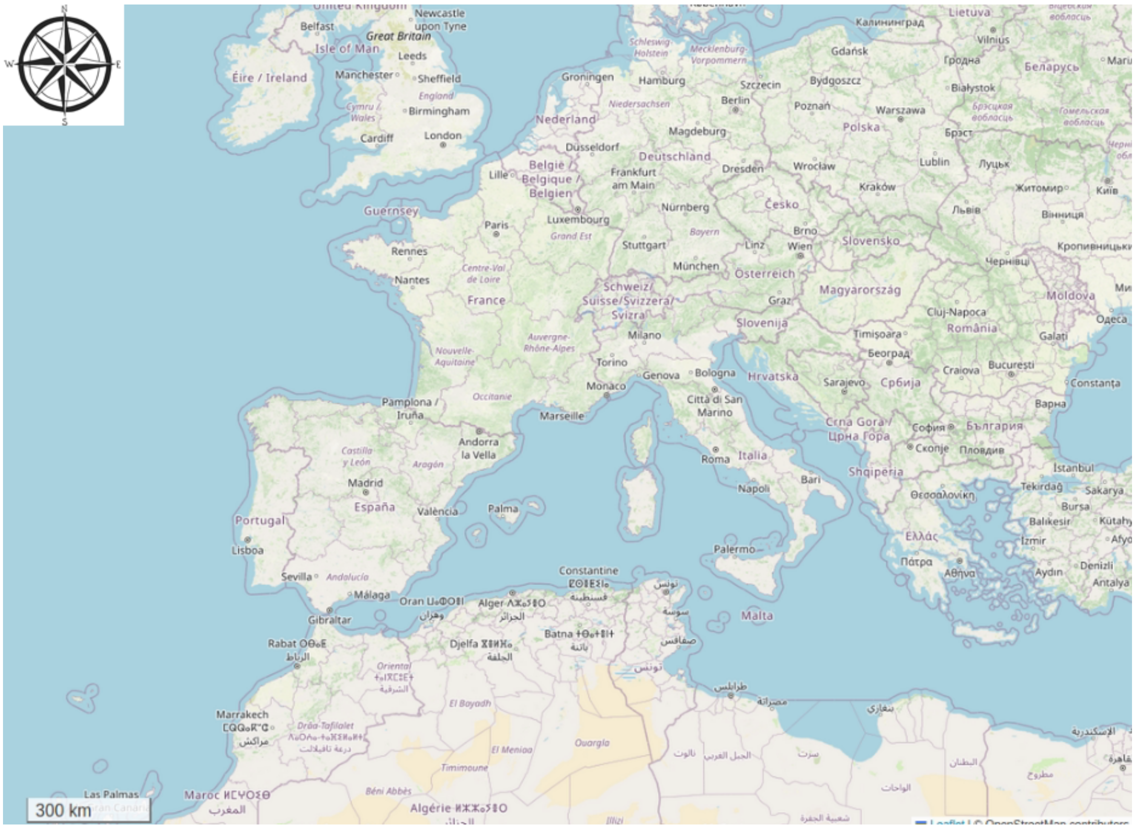
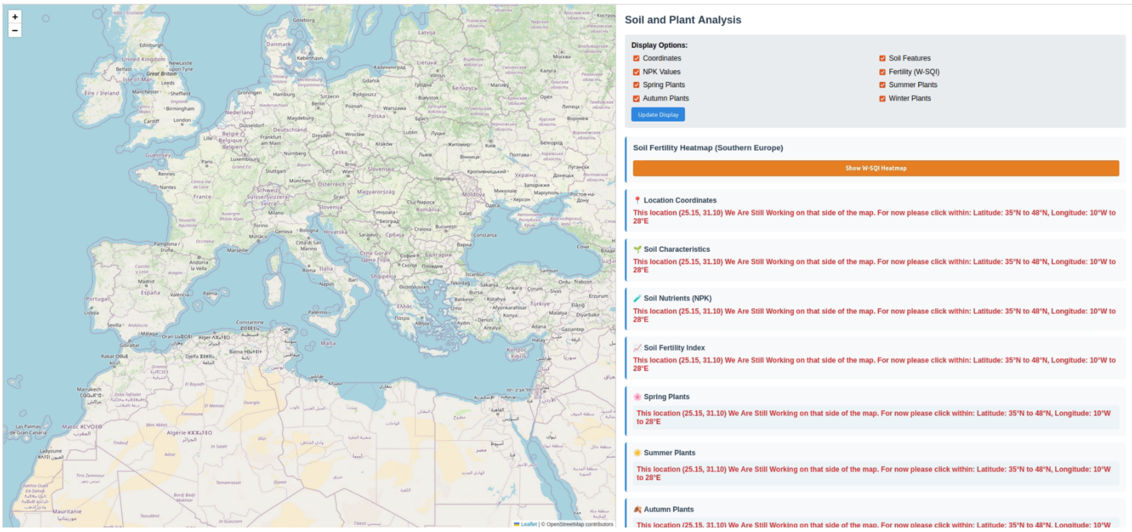


Figure 3.14: Interactive SQI and plant-suggestion map implemented with Leaflet.

When a user clicks on a location outside our study area, this interface in the figure 3.15 appears to notify them that data is currently unavailable for that region.



Soil and Plant Analysis

Display Options:

☑ Coordinates
☑ NPK Values
☑ Spring Plants
☑ Autumn Plants
☑ Soil Features
☑ Fertility (W-SQI)
☑ Summer Plants
☑ Winter Plants

Soil Fertility Heatmap (Southern Europe)

📍 Location Coordinates

This location (25.15, 31.10) We Are Still Working on that side of the map. For now please click within: Latitude: 35°N to 48°N, Longitude: 10°W to 28°E

🌱 Soil Characteristics

This location (25.15, 31.10) We Are Still Working on that side of the map. For now please click within: Latitude: 35°N to 48°N, Longitude: 10°W to 28°E

🧪 Soil Nutrients (NPK)

This location (25.15, 31.10) We Are Still Working on that side of the map. For now please click within: Latitude: 35°N to 48°N, Longitude: 10°W to 28°E

📈 Soil Fertility Index

This location (25.15, 31.10) We Are Still Working on that side of the map. For now please click within: Latitude: 35°N to 48°N, Longitude: 10°W to 28°E

🌸 Spring Plants

This location (25.15, 31.10) We Are Still Working on that side of the map. For now please click within: Latitude: 35°N to 48°N, Longitude: 10°W to 28°E

☀️ Summer Plants

This location (25.15, 31.10) We Are Still Working on that side of the map. For now please click within: Latitude: 35°N to 48°N, Longitude: 10°W to 28°E

🍂 Autumn Plants

This location (25.15, 31.10) We Are Still Working on that side of the map. For now please click within: Latitude: 35°N to 48°N, Longitude: 10°W

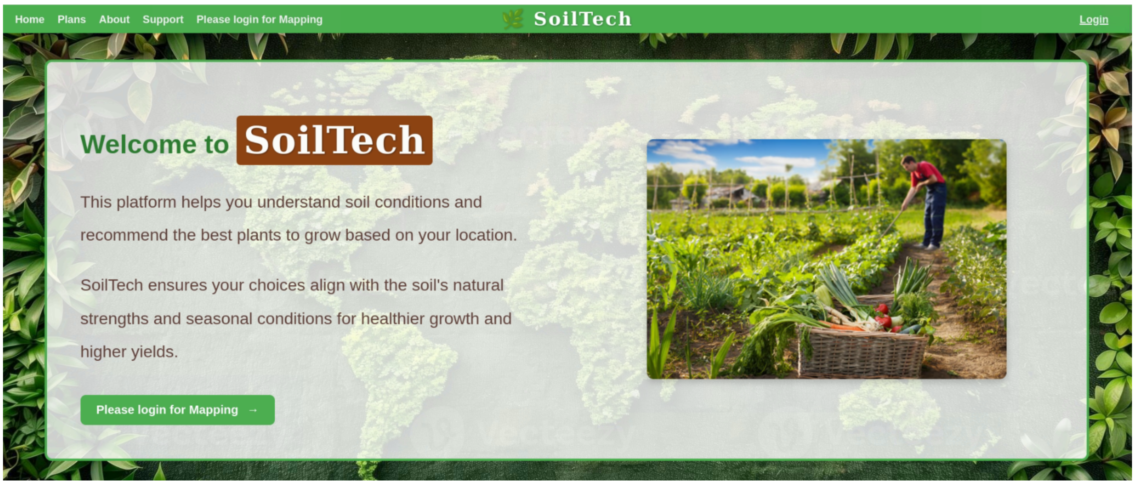
Figure 3.15: Alert shown when the selected point lies outside the study area (no data available).

3.4 SoilTech web application

Following the development of the prototype, we transitioned to a web-based application to ensure broader accessibility and practical use within the community. The web application was structured using standard web technologies, including HTML and CSS for the frontend interface design, and FastAPI for building the backend API services. This architecture enabled an interactive platform where users can explore predicted soil quality indicators and receive location-specific plant suggestion directly through a user-friendly interface. The web application was structured as follows:

Homepage

First you get to our application you find a homepage with an introductory section, plans section, our team section and the support contact as demonstrated in the figures 3.16



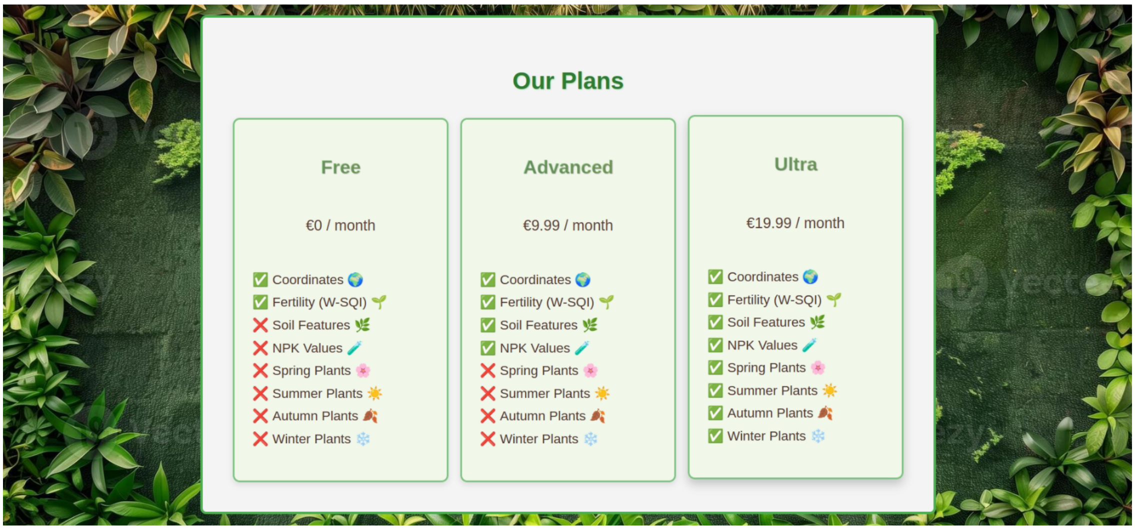
Welcome to **SoilTech**

This platform helps you understand soil conditions and recommend the best plants to grow based on your location.

SoilTech ensures your choices align with the soil's natural strengths and seasonal conditions for healthier growth and higher yields.

Figure 3.16: About us section

By scrolling down, the user will encounter the available plans section, as shown in Figure 3.17.



Our Plans

Free

€0 / month

- ✅ Coordinates 🌍
- ✅ Fertility (W-SQI) 🌱
- ❌ Soil Features 🌿
- ❌ NPK Values 🧪
- ❌ Spring Plants 🌸
- ❌ Summer Plants ☀️
- ❌ Autumn Plants 🍂
- ❌ Winter Plants ❄️

Advanced

€9.99 / month

- ✅ Coordinates 🌍
- ✅ Fertility (W-SQI) 🌱
- ✅ Soil Features 🌿
- ✅ NPK Values 🧪
- ❌ Spring Plants 🌸
- ❌ Summer Plants ☀️
- ❌ Autumn Plants 🍂
- ❌ Winter Plants ❄️

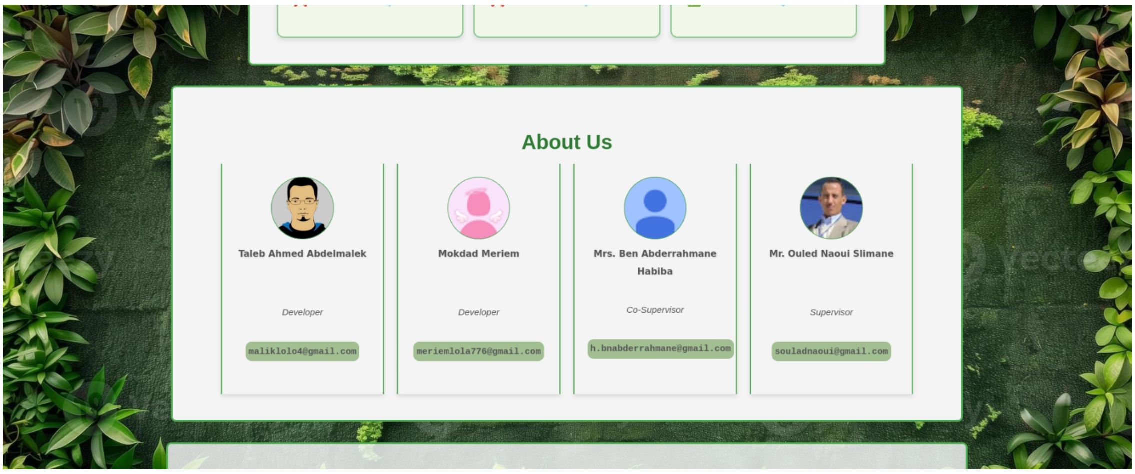
Ultra

€19.99 / month

- ✅ Coordinates 🌍
- ✅ Fertility (W-SQI) 🌱
- ✅ Soil Features 🌿
- ✅ NPK Values 🧪
- ✅ Spring Plants 🌸
- ✅ Summer Plants ☀️
- ✅ Autumn Plants 🍂
- ✅ Winter Plants ❄️

Figure 3.17: Plans section

Continuing downward, users are introduced to the team behind the application, showcasing the individuals involved and their contributions, as depicted in Figure 3.18.



About Us

Taleb Ahmed Abdelmalek
Developer
malikl olo4@gmail.com

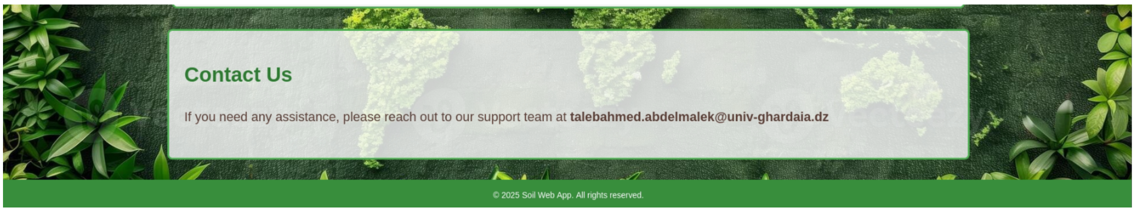
Mokdad Meriem
Developer
meriemlola776@gmail.com

Mrs. Ben Abderrahmane Habiba
Co-Supervisor
h.bnabderrahmane@gmail.com

Mr. Ouled Naoui Slimane
Supervisor
souladnaoui@gmail.com

Figure 3.18: Team section

Finally, the homepage concludes with the support contact section, offering users a way to reach out for assistance or inquiries. This section, shown in Figure 3.19, includes the email address of one of our team members for direct communication



Contact Us

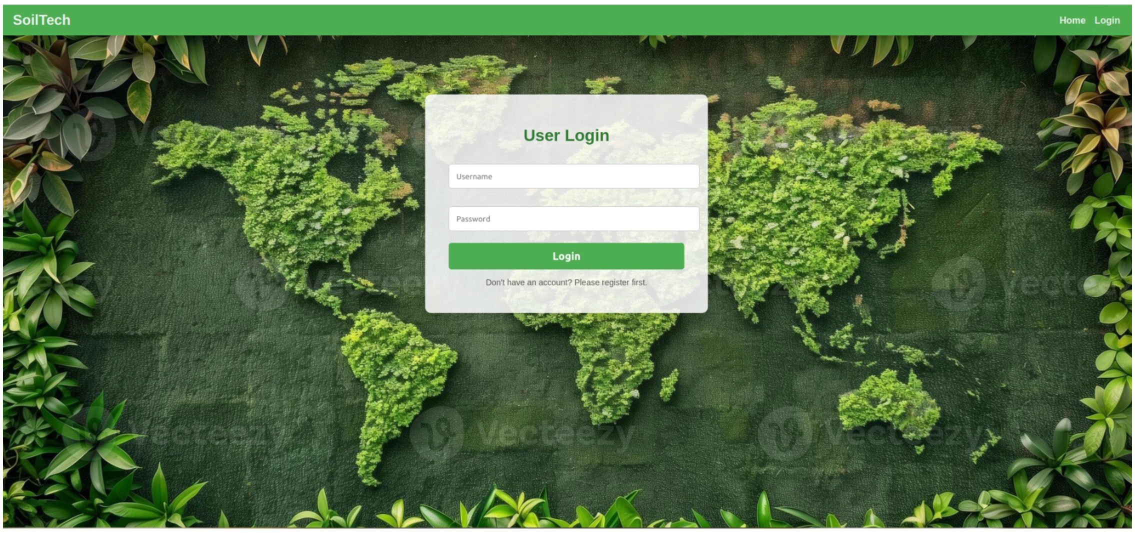
If you need any assistance, please reach out to our support team at **talebahmed.abdelmalek@univ-ghardaia.dz**

© 2025 Soil Web App. All rights reserved.

Figure 3.19: Support contact section

Login

To access the system's features, users must first log in through a dedicated interface. The login page is designed for simplicity and ease of use, allowing users to enter their credentials securely. This interface is illustrated in Figure 3.20.



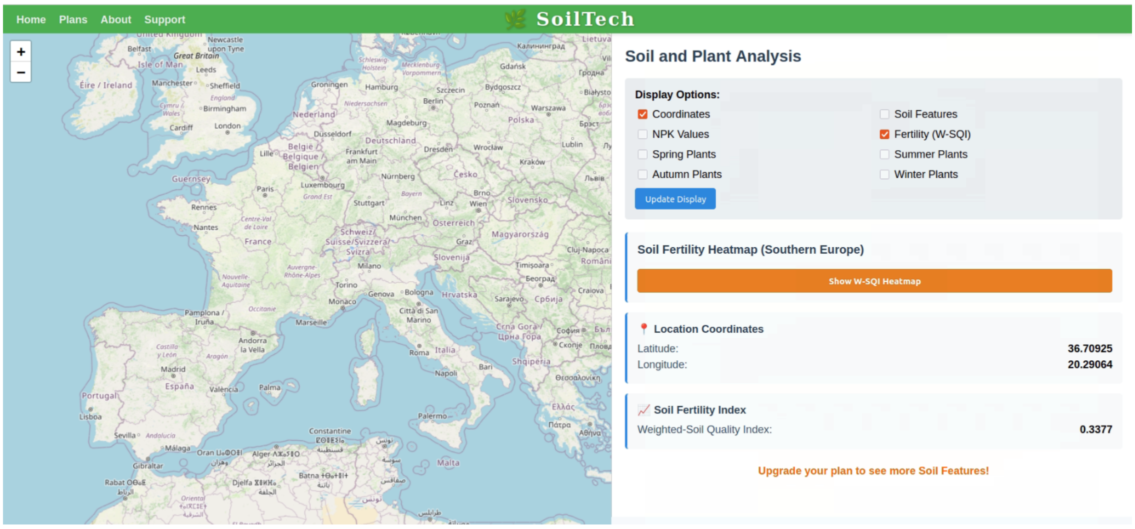
Don't have an account? Please register first.

Figure 3.20: Login section

System interactive map

Depending on the user's subscription plan, three distinct interfaces are available for the interactive map:

Free Plan Free-plan users have access only to the basic Soil Fertility Index (W-SQI) overlay.



Soil and Plant Analysis

Display Options:

☑ Coordinates ☐ Soil Features
☐ NPK Values ☑ Fertility (W-SQI)
☐ Spring Plants ☐ Summer Plants
☐ Autumn Plants ☐ Winter Plants

Soil Fertility Heatmap (Southern Europe)

📍 Location Coordinates

Latitude:	36.70925
Longitude:	20.29064

📈 Soil Fertility Index

Weighted-Soil Quality Index:	0.3377

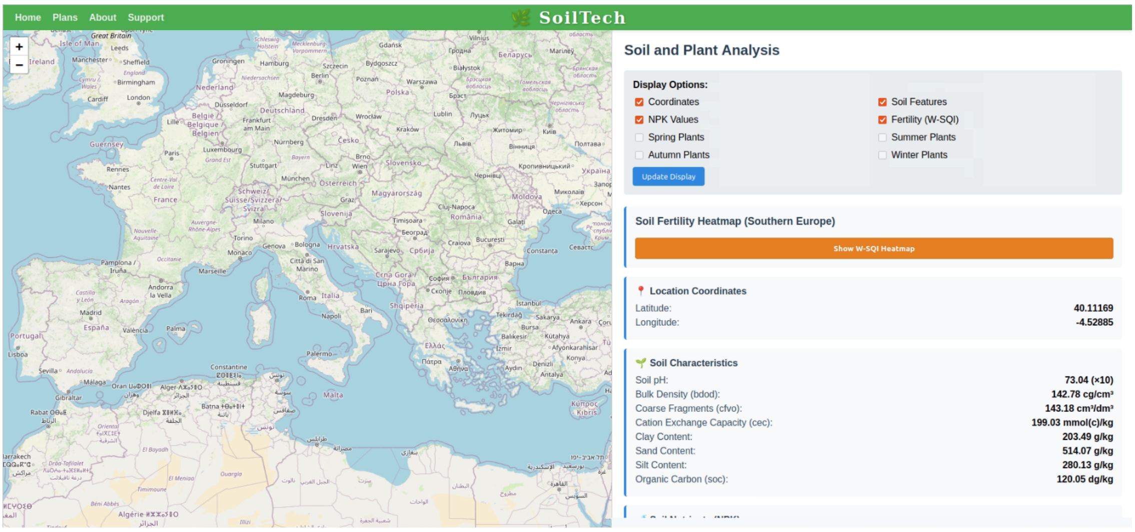
Upgrade your plan to see more Soil Features!

Figure 3.21: Free plan: interactive map displaying only the W-SQI heatmap.

Advanced Plan

Advanced-plan users can view all features except the seasonal plant suggestion.

Figure 3.22 showcases the Advanced plan, which includes an interactive map with access to soil parameters, NPK values, W-SQI, and a limited set of plant suggestions. In contrast, Figure 3.21 displays the Free plan, where users can only view the W-SQI heatmap without any additional data or recommendations.



Soil and Plant Analysis

Display Options:

☑ Coordinates
☑ NPK Values
☐ Spring Plants
☐ Autumn Plants
☑ Soil Features
☑ Fertility (W-SQI)
☐ Summer Plants
☐ Winter Plants

Soil Fertility Heatmap (Southern Europe)

📍 **Location Coordinates**

Latitude:	**40.11169**
Longitude:	**-4.52885**

🌱 **Soil Characteristics**

Soil pH:	**73.04 (×10)**
Bulk Density (bdod):	**142.78 cg/cm^3**
Coarse Fragments (cfvo):	**143.18 cm^3/dm^3**
Cation Exchange Capacity (cec):	**199.03 mmol(c)/kg**
Clay Content:	**203.49 g/kg**
Sand Content:	**514.07 g/kg**
Silt Content:	**280.13 g/kg**
Organic Carbon (soc):	**120.05 dg/kg**

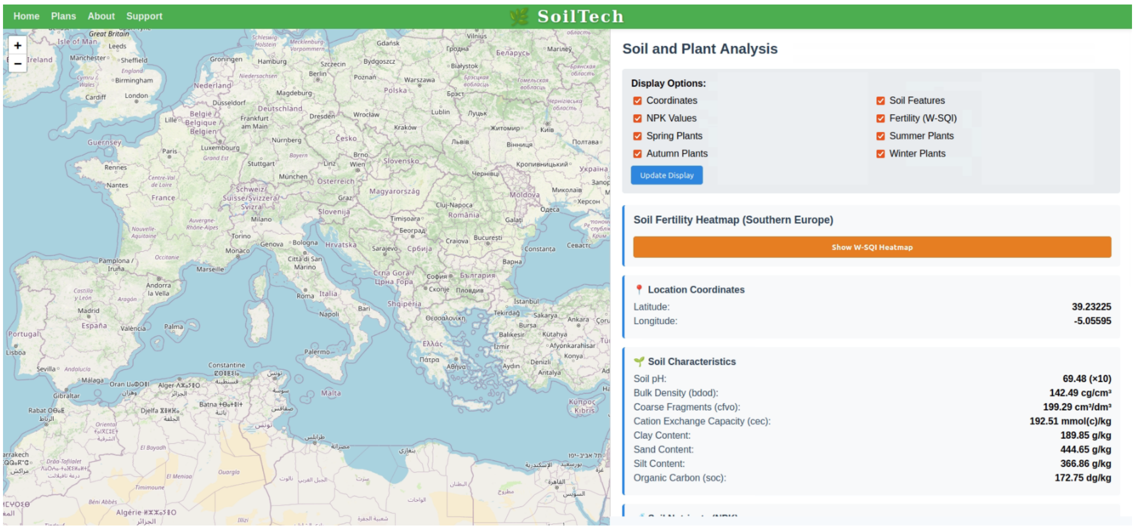
Figure 3.22: Advanced plan: interactive map with coordinates, soil parameters, NPK, W-SQI, and limited plant suggestions.

Ultra Plan

Users on the Ultra plan have full access to all map features, including:

- Display of geographic coordinates.
- Detailed soil characteristics (pH, bulk density, clay, sand, silt, organic carbon, cation exchange capacity, etc.).
- NPK nutrient values.
- Soil Fertility Index (W-SQI) heatmap.
- Top-three plant recommendations based on cosine similarity.

Figures 3.23 and 3.24 illustrate the main features of the Ultra plan interface. Figure 3.23 shows the fully interactive map, while Figure 3.24 presents an example of the detailed popup, which includes spatial coordinates, soil characteristics, the W-SQI, and tailored plant suggestions.



Soil and Plant Analysis

Display Options:

☑ Coordinates
☑ NPK Values
☑ Spring Plants
☑ Autumn Plants
☑ Soil Features
☑ Fertility (W-SQI)
☑ Summer Plants
☑ Winter Plants

Soil Fertility Heatmap (Southern Europe)

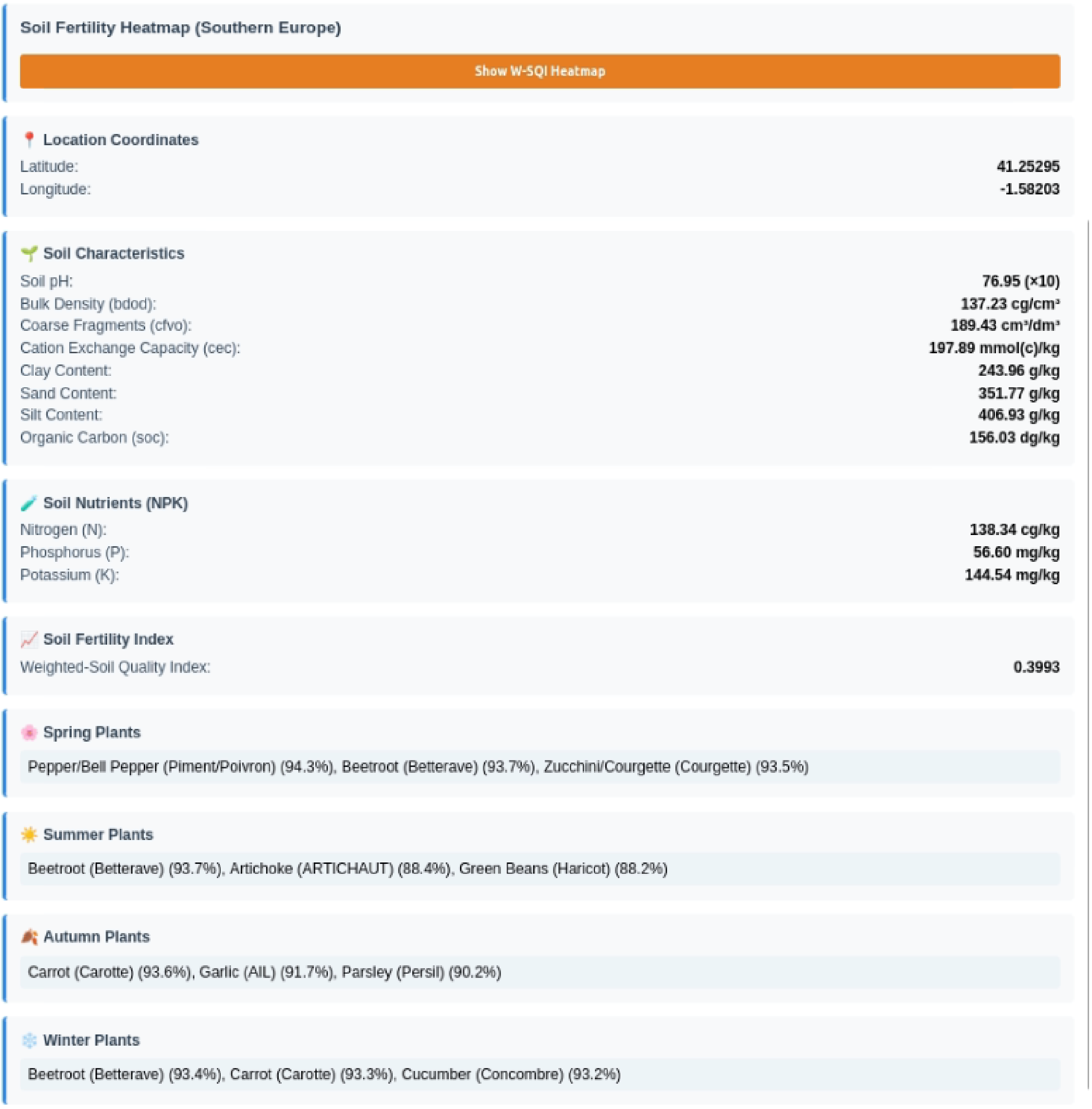
📍 **Location Coordinates**

Latitude:	**39.23225**
Longitude:	**-5.05595**

🌱 **Soil Characteristics**

Soil pH:	**69.48 (×10)**
Bulk Density (bdod):	**142.49 cg/cm^3**
Coarse Fragments (cfvo):	**199.29 cm^3/dm^3**
Cation Exchange Capacity (cec):	**192.51 mmol(c)/kg**
Clay Content:	**189.85 g/kg**
Sand Content:	**444.65 g/kg**
Silt Content:	**366.86 g/kg**
Organic Carbon (soc):	**172.75 dg/kg**

Figure 3.23: Ultra plan: full interactive map interface.



Soil Fertility Heatmap (Southern Europe)

📍 Location Coordinates

Latitude:	41.25295
Longitude:	-1.58203

🌱 Soil Characteristics

Soil pH:	76.95 (×10)
Bulk Density (bdod):	137.23 cg/cm^3
Coarse Fragments (cfvo):	189.43 cm^3/dm^3
Cation Exchange Capacity (cec):	197.89 mmol(c)/kg
Clay Content:	243.96 g/kg
Sand Content:	351.77 g/kg
Silt Content:	406.93 g/kg
Organic Carbon (soc):	156.03 dg/kg

🧪 Soil Nutrients (NPK)

Nitrogen (N):	138.34 cg/kg
Phosphorus (P):	56.60 mg/kg
Potassium (K):	144.54 mg/kg

📈 Soil Fertility Index

Weighted-Soil Quality Index:	0.3993

🌸 Spring Plants

Pepper/Bell Pepper (Piment/Poivron) (94.3%), Beetroot (Betterave) (93.7%), Zucchini/Courgette (Courgette) (93.5%)

☀️ Summer Plants

Beetroot (Betterave) (93.7%), Artichoke (ARTICHAUT) (88.4%), Green Beans (Haricot) (88.2%)

🍂 Autumn Plants

Carrot (Carotte) (93.6%), Garlic (AIL) (91.7%), Parsley (Persil) (90.2%)

❄️ Winter Plants

Beetroot (Betterave) (93.4%), Carrot (Carotte) (93.3%), Cucumber (Concombre) (93.2%)

Figure 3.24: Ultra plan: example popup showing coordinates, soil data, W-SQI, and plant suggestions.

3.5 Conclusion

In this chapter, we outlined two complementary experiments constituting the core of our Soil Quality Prediction (SQP) system. First, we compared four regression algorithms—LightGBM, XGBoost, Radial Basis Function Network, and Deep Neural Network—and selected XGBoost as the most accurate and efficient model to map the Weighted Soil Quality Index (W-SQI) across our study area. A Folium interactive map was then utilized to visualize spatial SQI predictions so that stakeholders could identify high and low soil fertility zones.

Second, we widened the pipeline to provide data-driven plant recommendation by establishing cosine similarity between predicted site feature vectors and expert-derived plant requirement norms. This stage was finalized with an interactive Leaflet map that provides personalized plant recommendations, soil characteristics, and SQI values at user-selected locations. Access to premium layers and recommendation features is managed via subscription levels, ranging from free-tier basic SQI overlays to full Ultra-plan functionality.

Together, these experiments demonstrate the promise and utility of combining machine learning, spatial analysis, and web-based visualization to make data-driven agricultural planning decisions.

Conclusion and Perspectives

This work outlines a machine learning-based approach for soil quality prediction and plant suggestion, divided into two basic experiments. The first experiment tries to evaluate the performance of four ML models—RBFN, LightGBM, XGBoost, and DNN—on a geospatial dataset of nine leading soil features. XGBoost achieved the best results, with an R^2 of 0.98, confirming its suitability for SQP tasks. The second experiment devises a two-stage framework: stage one consists of 36 expert regressors, where each model predicts a specific soil or environmental parameter from spatial coordinates. The predicted features are fed as input to a Random Forest Regressor for SQI prediction describing the soil fertility. Stage two contains a cosine similarity-based algorithm for matching predicted environmental vectors with species-specific requirements to suggest suitable plants. The proposed system is embedded in an interactive web application that offers visualization through maps, performs SQI analysis, and gives suggestions on plant species.

Future research will address current limitations by prioritizing the collection of more data, especially from Algeria, and extending the interactive map to greater geographical expanses, thereby enhancing the tool's usefulness in precision agriculture and environmental planning.

Bibliography

Bishop, C. M. and Bishop, H. (2024). *Deep Learning: Foundations and Concepts.* Springer International Publishing, Cham.

Brady, N. C. and Weil, R. R. (2016). *The Nature and Properties of Soils.* Pearson Education, Upper Saddle River, NJ, 15th edition.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Breure, T., Prout, J., Haefele, S., Milne, A., Hannam, J., Moreno-Rojas, S., and Corstanje, R. (2022). Comparing the effect of different sample conditions and spectral libraries on the prediction accuracy of soil properties from near- and mid-infrared spectra at the field-scale. *Soil and Tillage Research*, 215:105196.

Chang, K.-T. (2019). *Introduction to geographic information systems.* McGraw-Hill Education, New York, ninth edition edition.

Damiba, W. A. F., Gathenya, J. M., Raude, J. M., and Home, P. G. (2024). Soil quality index (SQI) for evaluating the sustainability status of Kakia-Esamburmbur catchment under three different land use types in Narok County, Kenya. *Heliyon*, 10(5):e25611.

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. In *Advances in Geophysics*, volume 61, pages 1–55. Elsevier.

Du, Z., Hu, Y., Wu, W., Lu, Y., and Buttar, N. A. (2020). Structural analysis on cutting notch of tea stalk by X-ray micro-computed tomography. *Information Processing in Agriculture*, 7(2):242–248.

El Behairy, R. A., El Arwash, H. M., El Baroudy, A. A., Ibrahim, M. M., Mohamed, E. S., Kucher, D. E., and Shokr, M. S. (2024a). How Can Soil Quality Be Accurately and Quickly Studied? A Review. *Agronomy*, 14(8):1682.

El Behairy, R. A., El Arwash, H. M., El Baroudy, A. A., Ibrahim, M. M., Mohamed, E. S., Rebouh, N. Y., and Shokr, M. S. (2024b). An accurate approach for predicting soil quality based on machine learning in drylands. *Agriculture (Nitra, Slovakia)*, 14(4):627.

Flach, P. A. (2012). *Machine learning: the art and science of algorithms that make sense of data.* Cambridge University Press, Cambridge ; New York. OCLC: ocn795181906.

Folorunso, O., Ojo, O., Busari, M., Adebayo, M., Joshua, A., Folorunso, D., Ugwunna, C. O., Olabanjo, O., and Olabanjo, O. (2023). Exploring machine learning models for soil nutrient properties prediction: A systematic review. *Big Data Cogn. Comput.*, 7(2):113.

Guerraoui, R., Gupta, N., and Pinot, R. (2024). Basics of machine learning. In *Robust Machine Learning: Distributed Methods for Safe AI*, pages 15–31. Springer Nature Singapore, Singapore.

Hemmati-Sarapardeh, A., Larestani, A., Nait Amar, M., and Hajirezaie, S. (2020). Chapter 3 - training and optimization algorithms. In Hemmati-Sarapardeh, A., Larestani, A., Nait Amar, M., and Hajirezaie, S., editors, *Applications of Artificial Intelligence Techniques in the Petroleum Industry*, pages 51–78. Gulf Professional Publishing.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, 15(14):5481–5487.

Huang, J.-C., Ko, K.-M., Shu, M.-H., and Hsu, B.-M. (2020). Application and comparison of several machine learning algorithms and their integration models in regression problems. *Neural Computing and Applications*, 32(10):5461–5469.

Inazumi, S., Intui, S., Jotisankasa, A., Chaiprakaikeow, S., and Kojima, K. (2020). Artificial intelligence system for supporting soil classification. *Results in Engineering*, 8:100188.

King, T. S., Chinchilli, V. M., and Carrasco, J. L. (2007). A repeated measures concordance correlation coefficient. *Statistics in Medicine*, 26(16):3095–3113.

Korstanje, J. (2023). *Machine learning on geographical data using Python: introduction into geodata with applications and use cases*. Apress, New York, NY.

Kumar, S. and Bhatnagar, V. (2022). A Review of Regression Models in Machine Learning. *JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTING*, 3(1):40–47.

Kursa, M. B. and Rudnicki, W. R. (2010). Feature Selection with the **Boruta** Package. *Journal of Statistical Software*, 36(11).

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lin, G.-F. and Chen, L.-H. (2004). A spatial interpolation method based on radial basis function networks incorporating a semivariogram model. *Journal of Hydrology*, 288(3-4):288–298.

Onyutha, C. (2020). From R-squared to coefficient of model accuracy for assessing "goodness-of-fits".

Padarian, J., Minasny, B., and McBratney, A. (2019). Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional*, 16:e00198.

Peng, Y., Liu, Z., Lin, C., Hu, Y., Zhao, L., Zou, R., Wen, Y., and Mao, X. (2022). A new method for estimating soil fertility using extreme gradient boosting and a backpropagation neural network. *Remote Sensing*, 14(14):3311.

Pham, V., Weindorf, D. C., and Dang, T. (2021). Soil profile analysis using interactive visualizations, machine learning, and deep learning. *Computers and Electronics in Agriculture*, 191:106539.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7(1):217–240.

Ramezanizadeh, M., Ahmadi, M. H., Nazari, M. A., Sadeghzadeh, M., and Chen, L. (2019). A review on the utilized machine learning approaches for modeling the dynamic viscosity of nanofluids. *Renewable and Sustainable Energy Reviews*, 114:109345.

Rogers, S. R., Foust, B. G., and Nelson, J. R. (2024). Standard use of Geographic Information System (GIS) techniques in honey bee research 2.0. *Journal of Apicultural Research*, pages 1–90.

Sepehya, S., Mehta, D., Kumar, A., Sharma, R., Sharma, D., and Sharma, A. (2024). Concept and Assessment Methodology of Soil Quality: A Review. *International Journal of Plant & Soil Science*, 36(5):164–172.

Shields, M. D. and Zhang, J. (2016). The generalization of Latin hypercube sampling. *Reliability Engineering & System Safety*, 148:96–108.

Silvero, N. E., Demattê, J. A., Vieira, J. D. S., Mello, F. A. D. O., Amorim, M. T. A., Poppiel, R. R., Mendes, W. D. S., and Bonfatti, B. R. (2021). Soil property maps with satellite images at multiple scales and its impact on management and classification. *Geoderma*, 397:115089.

Su, W., Jiang, F., Shi, C., Wu, D., Liu, L., Li, S., Yuan, Y., and Shi, J. (2023). An XGBoost-Based Knowledge Tracing Model. *International Journal of Computational Intelligence Systems*, 16(1):13.

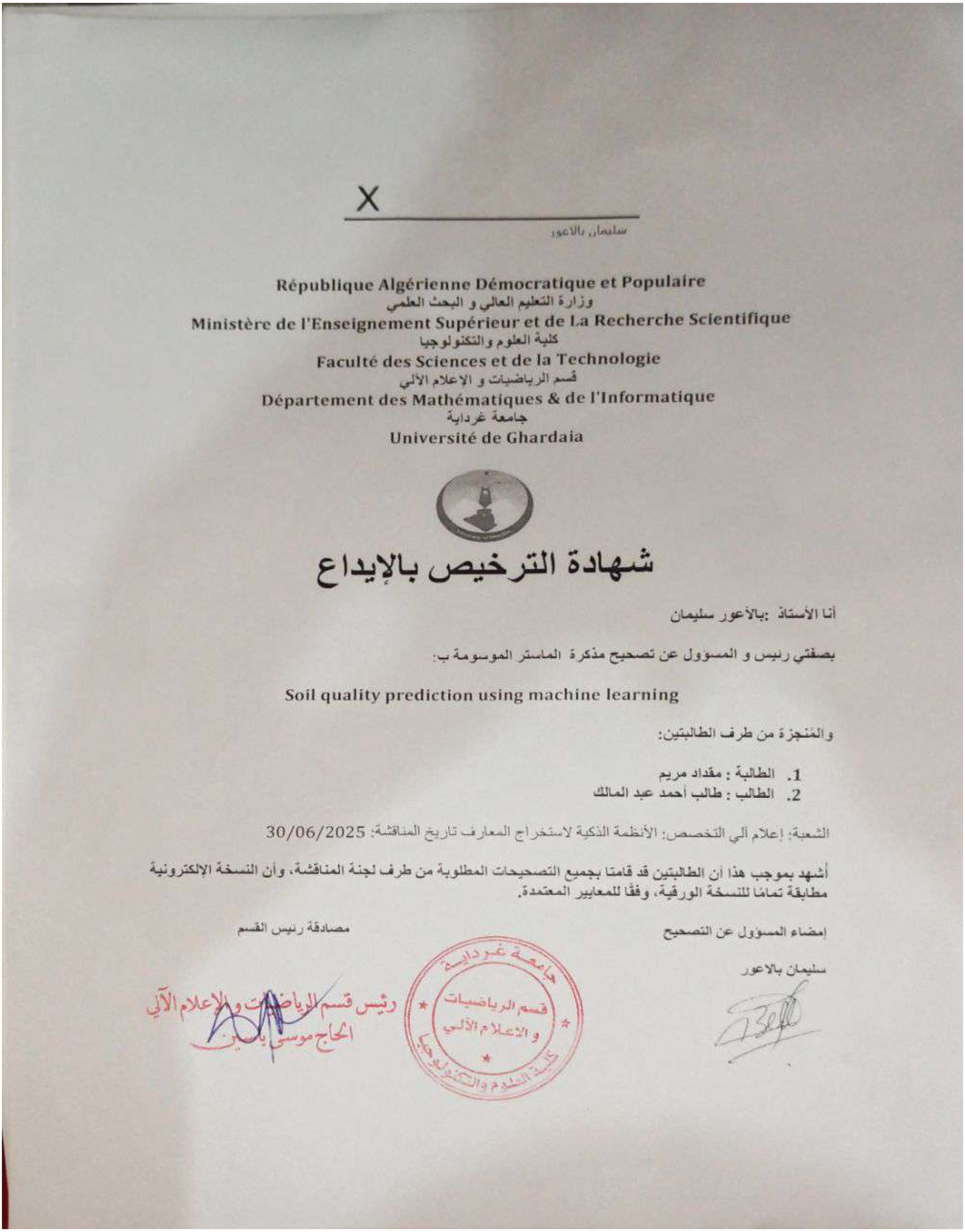
Sumathi, P., V. Karthikeyan, V., S. Kavitha, M., and Karthik, S. (2023). Improved Soil Quality Prediction Model Using Deep Learning for Smart Agriculture Systems. *Computer Systems Science and Engineering*, 45(2):1545–1559.

Yang, P., Zhou, J., Zhang, Y., Xu, C., Khandelwal, M., and Huang, S. (2025). Ground Settlement Prediction in Urban Tunnelling: Leveraging Metaheuristic-Optimized Random Forest Models. *Arabian Journal for Science and Engineering*.

Zolfaghari Nia, M., Moradi, M., Moradi, G., and Taghizadeh-Mehrjardi, R. (2022). Machine learning models for prediction of soil properties in the riparian forests. *Land*, 12(1):32.

APPENDIX A

ANNEX : DEPOSIT LICENSE CERTIFICATE



X

سليمان بالاعور

République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de La Recherche Scientifique
كلية العلوم والتكنولوجيا
Faculté des Sciences et de la Technologie
قسم الرياضيات و الإعلام الآلي
Département des Mathématiques & de l'Informatique
جامعة غرداية
Université de Ghardaia

شهادة الترخيص بالإيداع

أنا الأستاذ :بالأعور سليمان

بصفتي رئيس و المسؤول عن تصحيح مذكرة الماستر الموسومة ب:

Soil quality prediction using machine learning

والمنجزة من طرف الطالبتين:

1. الطالبة : مقداد مريم
2. الطالب : طالب أحمد عبد المالك

الشعبة: إعلام آلي التخصص: الأنظمة الذكية لاستخراج المعارف تاريخ المناقشة: 30/06/2025

أشهد بموجب هذا أن الطالبتين قد قامتا بجميع التصحيحات المطلوبة من طرف لجنة المناقشة، وأن النسخة الإلكترونية مطابقة تمامًا للنسخة الورقية، وفقًا للمعايير المعتمدة.

إمضاء المسؤول عن التصحيح

سليمان بالاعور

مصادقة رئيس القسم

رئيس قسم الرياضيات و الإعلام الآلي
الحاج موسى ياسين

جامعة غرداية
قسم الرياضيات و الاعلام الآلي
كلية العلوم والتكنولوجيا