




## Article

# Assessment of Soil Suitability Using Machine Learning in Arid and Semi-Arid Regions

Maryem Ismaili <sup>1,2,\*</sup>, Samira Krimissa <sup>1</sup>, Mustapha Namous <sup>1</sup> , Abdelaziz Htitiou <sup>1</sup> , Kamal Abdelrahman <sup>3</sup>, Mohammed S. Fnais <sup>3</sup>, Rachid Lhissou <sup>4</sup> , Hasna Eloudi <sup>5</sup>, Elhousna Faouzi <sup>1</sup> and Tarik Benabdellouahab <sup>2</sup>

<sup>1</sup> Data4Earth Laboratory, Department of Geology, Sultan Moulay Slimane University, Beni Mellal 23000, Morocco

<sup>2</sup> National Agronomic Research Institute, Rabat 10000, Morocco

<sup>3</sup> Department of Geology & Geophysics, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia

<sup>4</sup> Centre ETE, INRS, 490 rue de la Couronne, Québec, QC G1K 9A9, Canada

<sup>5</sup> Department of Geology, Ibn Zohr University, Agadir 80000, Morocco

\* Correspondence: maryem.ismaili17@gmail.com

**Abstract:** Increasing agricultural production is a major concern that aims to increase income, reduce hunger, and improve other measures of well-being. Recently, the prediction of soil-suitability has become a primary topic of rising concern among academics, policymakers, and socio-economic analysts to assess dynamics of the agricultural production. This work aims to use physico-chemical and remotely sensed phenological parameters to produce soil-suitability maps (SSM) based on Machine Learning (ML) Algorithms in a semi-arid and arid region. Towards this goal an inventory of 238 suitability points has been carried out in addition to 14 physico-chemical and 4 phenological parameters that have been used as inputs of machine-learning approaches which are five MLA prediction, namely RF, XgbTree, ANN, KNN and SVM. The results showed that phenological parameters were found to be the most influential in soil-suitability prediction. The validation of the Receiver Operating Characteristics (ROC) curve approach indicates an area under the curve and an AUC of more than 0.82 for all models. The best results were obtained using the XgbTree with an AUC = 0.97 in comparison to other MLA. Our findings demonstrate an excellent ability for ML models to predict the soil-suitability using physico-chemical and phenological parameters. The approach developed to map the soil-suitability is a valuable tool for sustainable agricultural development, and it can play an effective role in ensuring food security and conducting a land agriculture assessment.

**Keywords:** precision agriculture; sentinel-2; random forest; XgbTree; digital soil mapping; remote sensing; agricultural management



**Citation:** Ismaili, M.; Krimissa, S.; Namous, M.; Htitiou, A.; Abdelrahman, K.; Fnais, M.S.; Lhissou, R.; Eloudi, H.; Faouzi, E.; Benabdellouahab, T. Assessment of Soil Suitability Using Machine Learning in Arid and Semi-Arid Regions. *Agronomy* **2023**, *13*, 165. <https://doi.org/10.3390/agronomy13010165>

Academic Editors: Xiuliang Jin, Hao Yang, Zhenhai Li, Changping Huang and Dameng Yin

Received: 30 November 2022

Revised: 20 December 2022

Accepted: 29 December 2022

Published: 4 January 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

As the world's population is rapidly growing day by day, so too does the pressure to expand and intensify the use of agricultural land; additionally, putting a strain on natural resources to meet the rising demand for food and agricultural products [1–3]. This pressure can cause a degradation in the potential of agricultural lands, causing a list of issues, such as soil degradation, waterlogging, salinization/alkalization, and pollution [4], which have a direct impact on food production and food security [5–8]. In addition, agricultural land use is more demanding in terms of soil quality and performance when compared to other land uses. This requirement stems from the fact that not all soils can be used for agriculture, and not all crops can be grown successfully under given soil conditions, due to the differing crop nutrient requirements and the physico-chemical properties of soils [9]. For this purpose, researchers consider it of great importance to map land suitability [1,10,11] with the main goals of establishing the sustainable use of agricultural lands and assessing the soil potential for agricultural purposes [1,11]. The status of land suitability is determined

by intrinsic soil properties (e.g., parent materials, soil texture, organic matter, slope, and depth) as well as characteristics that can be affected by human management (e.g., drainage, irrigation, soil and water quality, soil fertility, and crop management) [1]. The new model and analysis approach has led to a number of high quality decision-making tools, which perform complex treatments using a large number of variables [12,13].

Many researches have applied new approaches to map land suitability, such as linear combination [13,14], simple limitation [15], fuzzy-logic modeling [16], artificial neural networks [17], remote sensing RS [18,19], and Machine learning ML. The major domains where ML algorithms are frequently used are disease detection, business intelligence, industry automation, and sentiment analysis [20]. Furthermore, ML methods have made significant contributions to soil degradation and landslide susceptibility mapping, ground subsidence and groundwater potentiality estimation [21,22], Multi-Criteria Decision Analysis (MCDA), Multi-Criteria Evaluation (MCE) [21–23], and Analytical Hierarchy Process (AHP) [24,25]. Despite some limitations of AHP, MCE and MCDA are still the most commonly applied methods for land evaluation, especially on a small scale. However, the AHP, MCE and MCDA methods are time-consuming and involve costly procedures in sampling and ground surveying without allowing to cover the spatio-temporal proprieties.

On this basis, ML and RS can address these limitations by offering an alternative means of classical soil suitability mapping in large spatio-temporal scales. ML models are capable of learning from large datasets and can integrate different types of data easily [26–28]. In a digital soil mapping framework, ML models were applied to make the links between soil observations and auxiliary variables in order to understand the spatial-temporal variation in soil types and other soil properties [29]. Additionally, the application of ML is well-known in assessing various phenomena in relation to natural disasters [30], soil erosion, and crop growth management [31]. Several algorithms were used to perform soil suitability maps based on artificial neural networks (ANNs) [30], random forest (RF) [26,32], support vector machines (SVMs) [31,33], K-Nearest Neighbor (K-NN) [34], and Extreme Gradient Boosting (XgbTree) [35]. Likewise, ML methods have demonstrated greater robustness and stability, making them popular and cost-effective in assessing agricultural land potentiality [27]. ML methods are still an emerging and challenging research field. In the present study, phenological characteristics derived from a vegetation indices time series were combined to fill the missing spatial information from the punctual soil observation data in order to assess for an agricultural land use potential. According to several studies and researches, the exploitation of phenological data has proven useful for detecting, mapping, and monitoring soil suitability [5,19]. The behavior of the vegetation cover expresses the potential of the soil (high or low) to produce biomass [18,35,36]. Using the maximum values of the phenological parameters, which are closely related to biomass production over a long period of time, other factors related to climatic conditions and crop management are directly excluded, and only the agricultural production potential of the land is highlighted [37].

The phenology approach indicates the soil suitability impact on vegetation development, which constitutes a key indicator for assessing soil potential. However, in order to assess for a land suitability map for agriculture, a deep understanding of soil type characteristics and vegetation cover behavior is highly required. In this case, the ML methods were selected to perform the spatial analysis, combining phenological metrics and physico-chemical soil factors. The uniqueness of this work is the development of a decisive, fast, efficient, and less expensive approach for evaluating soil suitability with the greatest precision. We have developed an original and novel method for this purpose that combines physico-chemical and phenological approaches using artificial intelligence tools. To our knowledge, this novel approach will be tested for the first time in the field of smart agriculture.

The main purpose of this study is to map soil suitability in a semi-arid and arid region, using five ML methods. In this regard, a total of eighteen physico-chemical soil factors (N, ESP, Capacity of Exchange Cation (CEC), Ca, P, K, pH, Na, OM, Mg, CaCO<sub>3</sub>, depth, and soil salinity) and four main phenological parameters were chosen to constitute a spatial

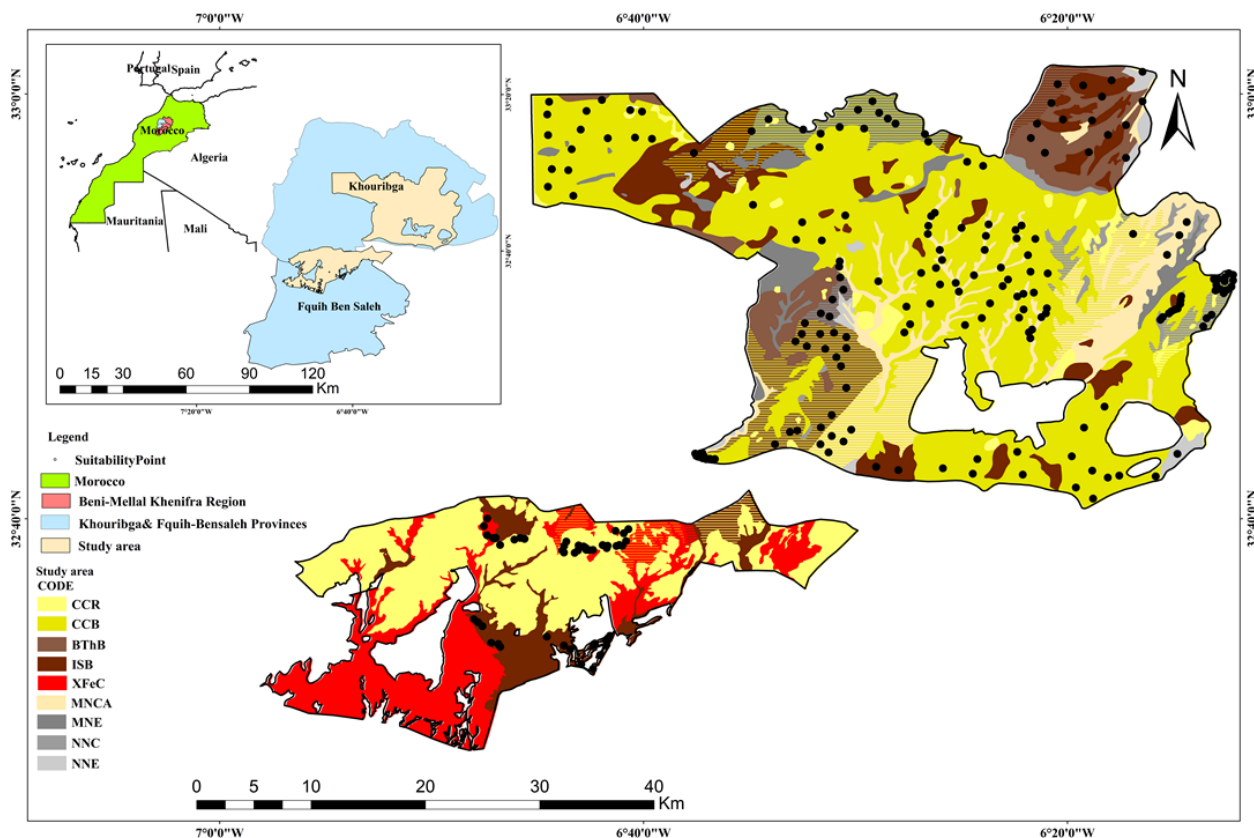
input to establish decisional maps based on ML. Consequently, this innovative approach can help systematically to assess the actual situation of soil suitability by considering the aforementioned factors. The combination of soil physico-chemical parameters and phenological factors related to biomass makes this work original and a useful tool for decision makers. The study's findings will assist decision makers and stakeholders in achieving sustainable development goals and continuing to increase agricultural production in the region.

## 2. Materials and Methods

### 2.1. Materials

#### 2.1.1. Study Area

The study area is located in the Beni-Mellal-Khenifra region of central Morocco between latitudes of  $32^{\circ}30'0''$  and  $33^{\circ}40'0''$  N and longitudes of  $6^{\circ}10'0''$  and  $7^{\circ}00'0''$  W. (Figure 1). It covers an area of 1541 km<sup>2</sup>. The region has a semi-arid to arid climate, with a dry season lasting from April to October and a rainy season extending from November to March [38] with an annual rainfall of 350 mm and evaporation of 1800 mm, [35]. The average annual temperature is 17 °C, with winter months averaging 3.5 °C and summer months averaging 38 °C. The region's primary lithological formations are Lias limestones and dolomites, alluvium, Triassic red clays, Paleozoic shales, sandstones, and quartzites. Furthermore, the appropriateness of the region's soils is one of the main critical production elements in agriculture and can have a direct impact on crop productivity [23,38]. (Figure 1, Table 1).



**Figure 1.** Location of the study area and soil type. BThB: Browned soils, CCB: Calci-magnesian Limestone browns soils, CCR: Calci-magnesianrendzines soils, ISB: isohumic soils, NNCA: Little evolved soils, MNE: Raw mineral soils, NNC: Little evolved soils Of colluvial contribution, NNE: Little evolved soils Erosion, XFeC: iron sesquioxide soils.

**Table 1.** Description of the soil unit codes.

Classes	Sub-Classes	Groups	Codes
Raw mineral soils	No climatic	Erosion	MNE
Little evolved soils	No climatic	Erosion	NNE
Little evolved soils	No climatic	Colluvial contribution	NNC
Calci-magnesian soils	Carbonates	Rendzines	CCR
Calci-magnesian soils	Carbonates	Limestone browns	CCB
iso-humic soils	In pedoclimate in the rainy season	Subtropical browns	ISB
Browned soils	Temperate-humid climates	Browns	BThB
iron sesquioxide soils	Fersiallitics	Low-leaching calcium reserve	XFeC
Little evolved soils	No Climatic	Collu-alluvial contribution	NNCA

### 2.1.2. Soil Data

In order to achieve our objectives, in addition to the field missions carried out, we used data collected and analyzed by the National Institute of Agronomy (INRA). Over the research region, a total of 28 soil profiles were collected and evenly dispersed. The laboratory of the IAV Hassan II, used methods recognized and adopted to the Moroccan soils for the pedological classification (CPCS 1967) [39,40] to conduct the geochemical analyses. The study area is characterized by 10 soil classes with 4% crude mineral soils, 15% poor soils, 43% calcium-magnesium soils, 16% iso-humic soils, 7% brown soils, 7%, and 8% iron sesquioxide soils and the rest are complex soils (Table 1, Figure 1).

### 2.1.3. Satellite Data

In this study, Sentinel-2 MSI data were used to extract phenological data over the study area. The Sentinel-2 contains the Multispectral Instrument (MSI), which has significantly different spectral response functions with 13 spectral bands, three distinct spatial resolution, and a five-day interval between revisits when two satellites are operating simultaneously [40]. The current research aims to understand the soil suitability by integrating satellite-phenological metrics. Therefore, all the available sentinel-2 data over the study area were acquired through the Google earth engine platform, covering the period from September 2016 to August 2020 (150 images).

## 2.2. Methodology

Figure 2 depicts an overview of the methodology used in this study, outlining the steps (pre-processing, compositing, smoothing, phenological information extraction, and classification) used in mapping the soil suitability based on phenological information data and soil parameters.

### 2.2.1. The Suitability Inventory Map (SIM)

The SIM is required for various predictive models to prepare the Soil Suitability Map [9]. To develop the SIM, the mapping of suitable areas was carried out using high-resolution Google Earth images as well as soil suitability maps derived from INRA. As results showed 238 suitability points were identified. Then, 70% of soil suitability points (166) were randomly selected for model training, and the remaining 30% (72) were used for models testing (Figure 2) [30,41].

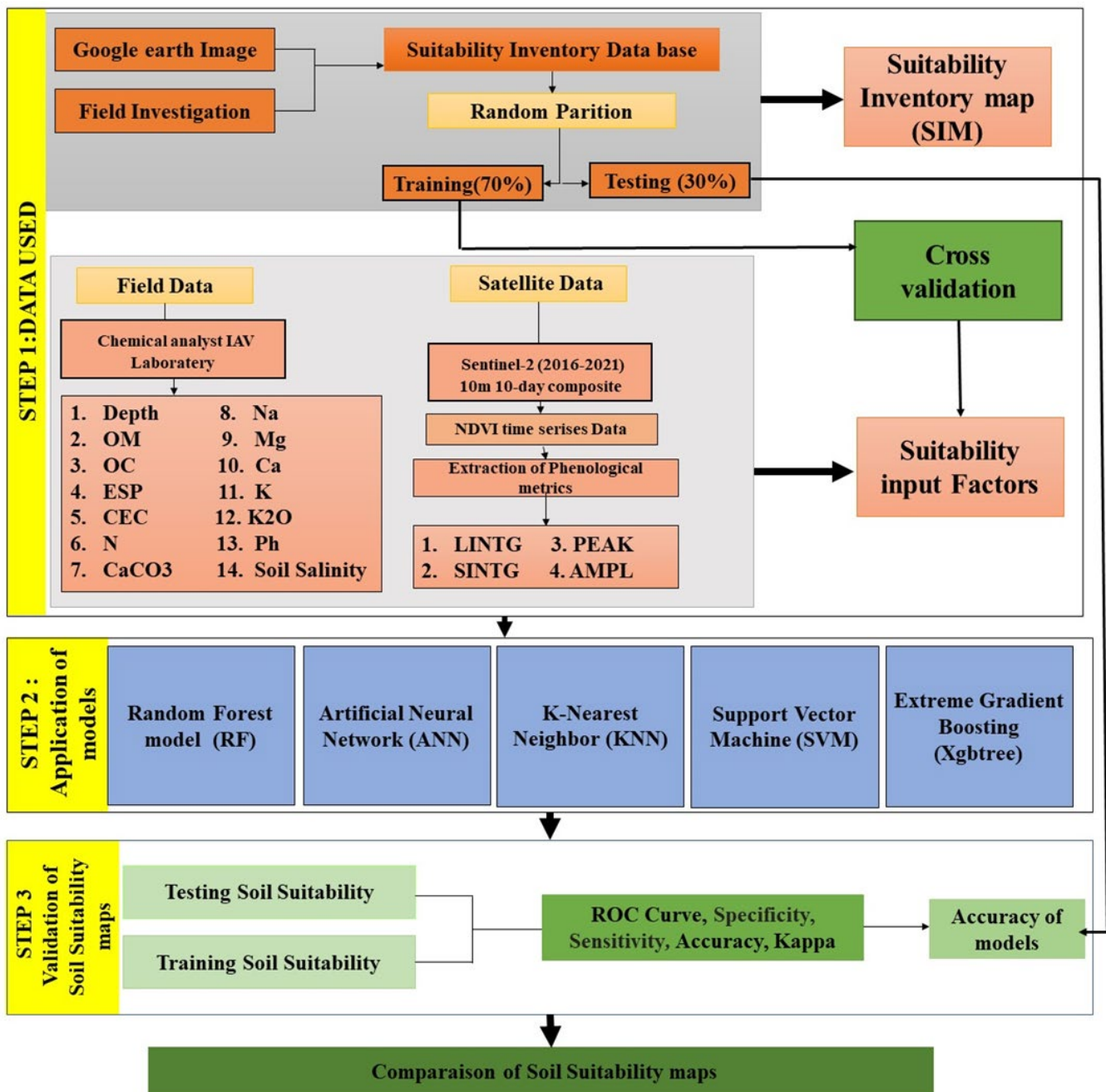


Figure 2. Flowchart of the methodology.

### 2.2.2. Processing of Geochemical Soil’s Parameters

In this work, point-based field data are utilized to generate a map of soil geochemical properties interpolated using the Inverse Distance Weighted (IDW) approach, and each measurement point is assigned a weight. The distance between the point and the other unsampled point influences the amount of weight applied. The power of ten is used to handle these weights. With a power-of-ten increase, the impact of the points that are farther away increases [42]. Weights are distributed across adjacent places more evenly when there is less power. The distance between the points matters in this strategy; therefore, points with the same distance have the same weights, Figure 3. After the IDW we classify the soil parameters according to the soil suitability classes [43]; we found that salinity is not a constraint affecting the soils in the study area.

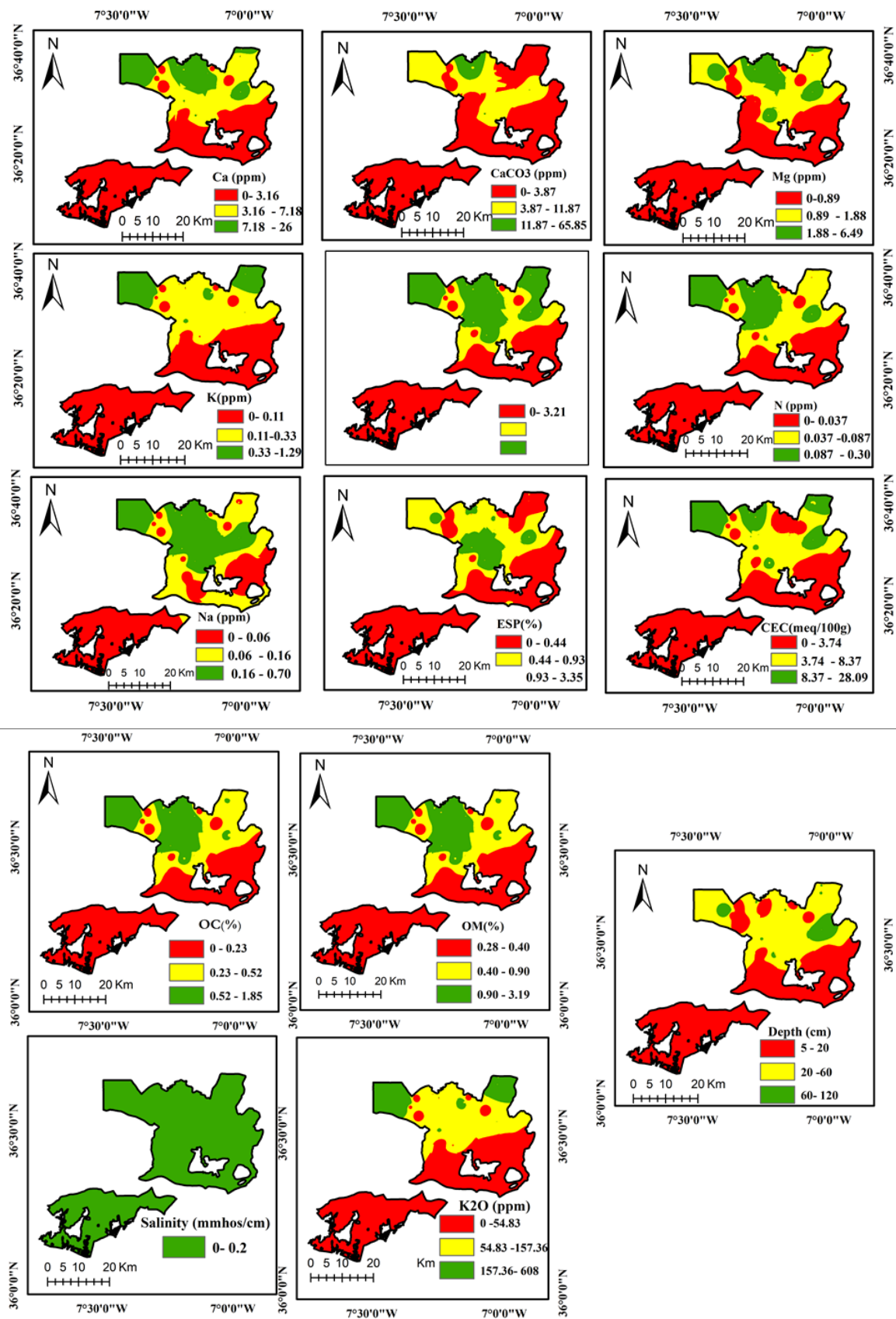


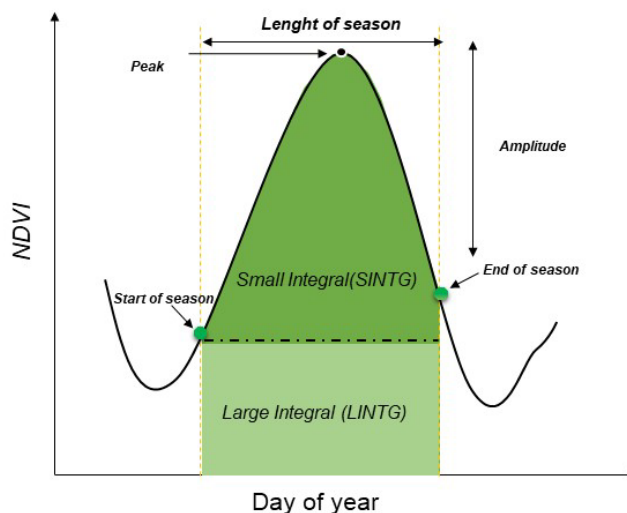
Figure 3. Physico-chemical Factors.

### 2.2.3. Processing Phenological Metrics

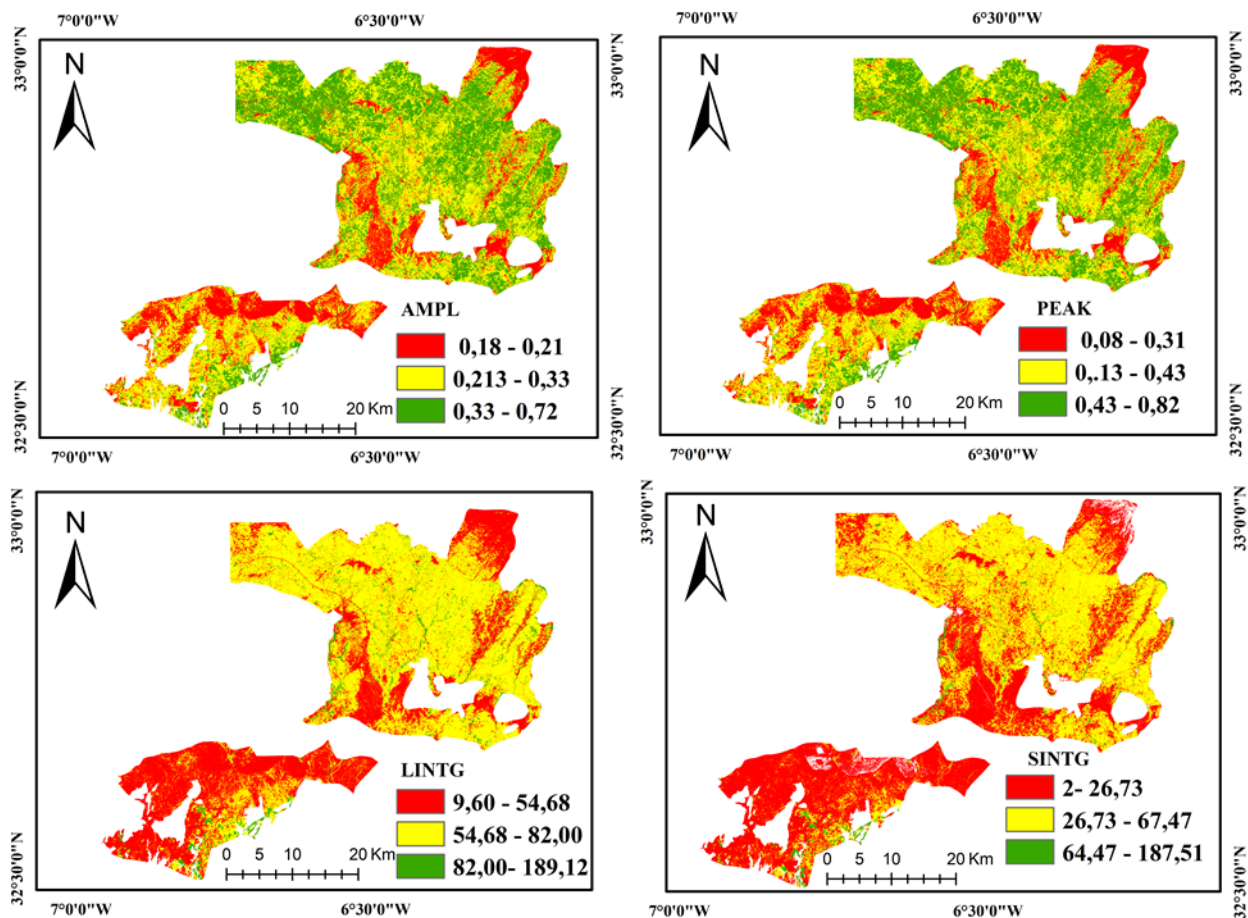
After generating a 10 day composite NDVI from the original Sentinel-2 data in Google Earth Engine (GEE), the Savitzky–Golay filter was applied to the NDVI profile to reduce the remaining noises in the time-series. The resulting filtered and fused NDVI time-series were then used to extract a set of 13 phenological metrics, which included the date of the start, mid, end, and length of the season and the phenological proxies, such as the peak and base values, seasonal amplitude or rate of increase and decrease (see the full list of metrics in Table 2) [43–49]. Several authors [35,45,46] have demonstrated that four phenological parameters (AMPL, PEAK, LINGT, and SINTG) are considered good indicators of biomass production. Overall, vegetation development expresses various factors (e.g., anthropic action, crop type, water deficiency, drought, pests, and nutrient deficiency) and not solely soil potential, hence the reason why the maximum values of phenological parameters were selected as input parameters. This approach allows for the exclusion of other factors related to climatic conditions and crop management, highlighting only the soil production potential (Figures 4 and 5).

**Table 2.** Seasonality parameters in TIMESAT.

Phenological Parameters	Abbreviation	Description
Start of season	SOS	Time for which the left edge has increased to 10% of the seasonal amplitude measured from the left minimum level.
End of season	EOS	Time for which the right edge has decreased to 10% of the seasonal amplitude measured from the right minimum level.
Middle of season	MOS	Mean value of the times for which the left part of the VGI curve has increased to the 90% level and the right part has decreased to the 90% level.
Length of season	LOS	Time from the start to the end of the season.
Base value	BVAL	The average of the left and right minimum values.
Maximum value	PEAK	Maximum VGI value for the fitted function during the season.
Amplitude	AMPL	Difference between the peak value and the base level.
Large integral	LINTG	The area under the smoothed curve between SOS and EOS.
Small integral	SINTG	The area below the base level from the SOS to EOS.
Left derivative	LDERIV	Rate of increase at the SOS between the left 10% and 90% of the amplitude.
Right derivative	RDERIV	Rate of decrease at the EOS between the right 10% and 90% of the amplitude.
Start of season value	SOSV	Start of season value.
End of season value	EOSV	End of season value.



**Figure 4.** Phenological parameters of the crop growing season used in this study.



**Figure 5.** Phenological Factors.

## 2.2.4. Machine Learning Algorithms

### K-Nearest Neighbor

The k-nearest neighbor algorithm, abbreviated KNN or k-NN, is one of the simplest machine learning algorithms based on the supervised learning technique. It assumes the new and current data are similar and places the new example in the category to which it is most similar. The KNN technique can be used for both regression and classification. The KNN algorithm is as follows: an accurate estimate of K is chosen, which should be an odd number, and the higher the K the greater the precision. The Euclidean distance between the test data point and all other data points is then calculated [30].

### Extreme Gradient Boosting Tree

Extreme Gradient Boosting (XgbTree) models use a scalable machine learning algorithm with an end-to-end tree boosting system to classify or predict the desired models from a larger learning database. It is a decision tree-based algorithm with gradient boosting as the core optimization technique. It belongs to the boosting family of algorithms where misclassification information from the previously grown tree is used to improve the next tree, making it an optimized sequential process, also known as 'boosting technique'. Tuning parameters is the hardest part of modeling because decision tree algorithms are known for overfitting. To obtain a model with low bias and low variance you have to find the best way to hyper-tune the parameters. This prevents the model from becoming too accurate on the training dataset, and it also enables the model to become more accurate on test datasets. With these features, the XgbTree model is a good choice for classifying soil and land crops that have almost identical spectral signatures when using multispectral data [49].



### Artificial Neural Network

In the Artificial Neural Network (ANN) model, the Multi-Layer Perceptron (MLP) architecture was chosen, which contains three layers connected by several neurons: the input layer, the hidden layer, and the output. For the input layer, which has one input and several output pathways for each neuron, each node is connected with different determining factors (GIFs). Hidden nodes, where there are several inputs and output connections for each neuron, use weighted connections to learn and process the problem; weights can take positive or negative values. Usually, for the modelling phase, the ANN method starts with the adjustment of the weights of the different connections between neurons during the training phase, then the output prediction stage is based on the constructed models [44].

### Support Vector Machine

Support Vector Machine (SVM) is a popular Supervised Learning algorithm that is used for both classification and regression problems. However, it is primarily used in machine learning for classification problems. The SVM algorithm's goal is to find the best line or decision boundary for categorizing n-dimensional space. In this case, a hyperplane is the best decision boundary, thus, SVM selects the extreme points/vectors that aid in the creation of this hyperplane. These extreme cases are referred to as support vectors, and the algorithm is known as the support vector machine [50].

### Random Forest

Random Forest (RF) Algorithm is an ensemble machine learning method used for both classification and regression. RF represent a group of decision trees in which the values of a random vector are selected separately and evenly throughout all trees in the forest to calculate the value of each tree. As the number of trees within a forest increases, the overfitting approaches a limit. The generalization error of a forest of tree classifiers is dependent on the relative strengths of the trees in the forest and their correlation. Using a random selection of features to split each node yields error rates. In order to control the inaccuracy, estimates are made of the model's response to an increase in the number of variables included in the analysis, as well as their relevance [50,51].

### Models Hyperparameters

For all models, we tested the hypertune to detect the hyperparameters of each model. The results are as follows: for RF with  $M_{try} = 10$  and  $M_{tree} = 500$ ; for KNN to choose the number of neighbors we applied  $K = 5$ ; for ANN with  $size = 5$  and  $decay = 5$ ; for SVM with  $sigma = 3$  and  $C\ index = 0.2$ ; and finally for `xgbtree`, the maximum number of iterations is 200, the tree depth is 5, the learning rate is 0.05, and the loss reduction is  $minimal = 0.05$ , minimum loss reduction required  $gamma = 0.01$   $colsample\_bytree = 0.75$ , and the sub-sample ratio of the training instance = 0.5.

### 2.2.5. Evaluation of the ML Algorithms Performance

#### Statistical Measures

To evaluate the robustness of the used machine learning models using the soil suitability modeling process we employed a number of statistically based metrics, including the kappa coefficient. The Cohen kappa method is used to evaluate a reliable soil suitability model. The value varies between +1 and -1. The result when close to 1 is the most suitable. The accuracy method generates the accurate model by combining suitable and non-suitable areas. To evaluate the SSA modeling process we use four types of possible outcomes—true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The suitable area and non-suitable area are obtained based on a cutoff value (here it is 0.5). Then, it is calculated based on the comparison between each suitability truth pixel and the suitability pixel on the obtained classified map. TP and FP refer to the suitability locations that are determined to be suitable and non-suitable locations. FN and TN classify non-suitability

locations as suitable and non-suitable locations, respectively. All of the equations used to calculate these parameters are mentioned below:

$$\text{Accuracy} = \frac{TN + TP}{TP + FP + TN + TP} \quad (1)$$

$$\text{Kappa} = \frac{\text{Accuracy} - B}{1 - B} \quad (2)$$

$$\text{where } B = \frac{(TP + FN)(TP + FP) + (FP + TN)(FN + TN)}{\sqrt{TP + TN + FN + FP}} \quad (3)$$

### ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve is considered a typical measure for assessing the outcomes of applying ML models [40]. The sensitivity and the 1-specificity combined with diverse cut-off limits are used to make the ROC curve. The perfect model has an area under the curve (AUC) value close to 1, whereas an inaccurate model is characterized by an AUC value of 0.5 [40]. In this study, the validation of the best-selected model was performed using the ROC curve

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (5)$$

## 3. Results

### 3.1. Relative Importance of the Factors Affecting Soil Suitability

In order to assess the power of phenological and environmental covariates to distinguish between suitable and non-suitable areas, we employed variable analysis to classify the most important factors, using RF, K-NN, ANN, SVM, and XgbTree methods. (Figure 6). The results of the RF model show that the most relevant factors to map soil suitability are  $\text{CaCO}_3$ ,  $\text{K}_2\text{O}$ , K, and CEC. However, the models K-NN, SVM, and XgbTree indicate that ESP, Na,  $\text{CaCO}_3$ , and the large integral are the most influential variables. In accordance with the ANN technique, the most relevant variables are ESP, Na, Mg, and  $\text{K}_2\text{O}$ . In general, both factors phenological and environmental covariates appear to be able to separate soil suitable from non-suitable. In particular, the productivity metrics, such as LINTG, SINTG, AMPL, PEAK, and the environmental covariates configure a good separability between the two classes. However, only a very small part of one particular class may be misclassified due to the existence of similar characteristics in phenological metrics that show less power discrimination among other features.

### 3.2. Spatial Soil Suitability Analysis

The soil suitability models were built using the five ML algorithms described above (Figure 7). The suitable area indices produced by the RF, ANN, XgbTree, KNN, and SVM models were classified into four classes according to the FAO, namely low, medium, high, and very high suitability, based on the natural break classification method. Figure 6 shows soil suitability maps produced by the five machine learning models. The SSM soil suitability map produced by the RF model (Figure 7) found that 30% of the study area has a very high soil suitability. The high and moderate SS zones cover 19% and 17% of the study area, respectively, while the remaining 35% falls into the low soil suitability class (Figure 7). The SSM soil suitability map produced by the ANN model (Figure 7) found that more than 54% of the study area has a low soil suitability. The high SS zones cover 3% and the moderate cover less than 3% of the study area, while the remaining 40% falls into the very high soil suitability class (Figure 7). Based on the results of XgbTree (Figure 7), the research area has 41% that falls into the low suitability class, followed by 40% in the very high suitability

class, 8.04% in the high suitability class, and 11% in the moderate SS classes (Figure 7). In case of the KNN model (Figure 7), the results show that the study area has a 25, 23% area in the very high suitability class, followed by the moderate and high class at 24.07% and 19.85% respectively, and 30.8% in the Low SS classes (Figure 7, Table 3). According to this model, ESP and Na are the most important contributing factors for MSS (Figure 6), while phenology plays an important role in the soil suitability characterization in the case of the RF and xgbtree models, as opposed to ANN, where the phenology plays a less important role. In the case of the SVM model (Figure 7), the results show that the study area has 38.35% area in the very high suitability class, followed by the low and high class at 30.63% and 17.45%, respectively, and 13.52% in the moderate SS classes (Figure 7).

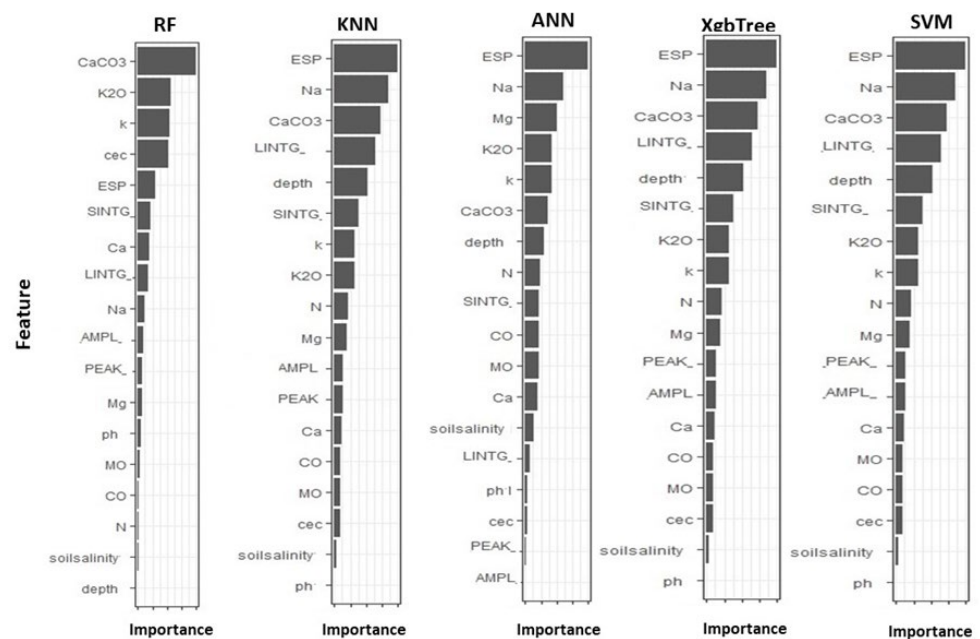


Figure 6. Factors of importance in the five models.

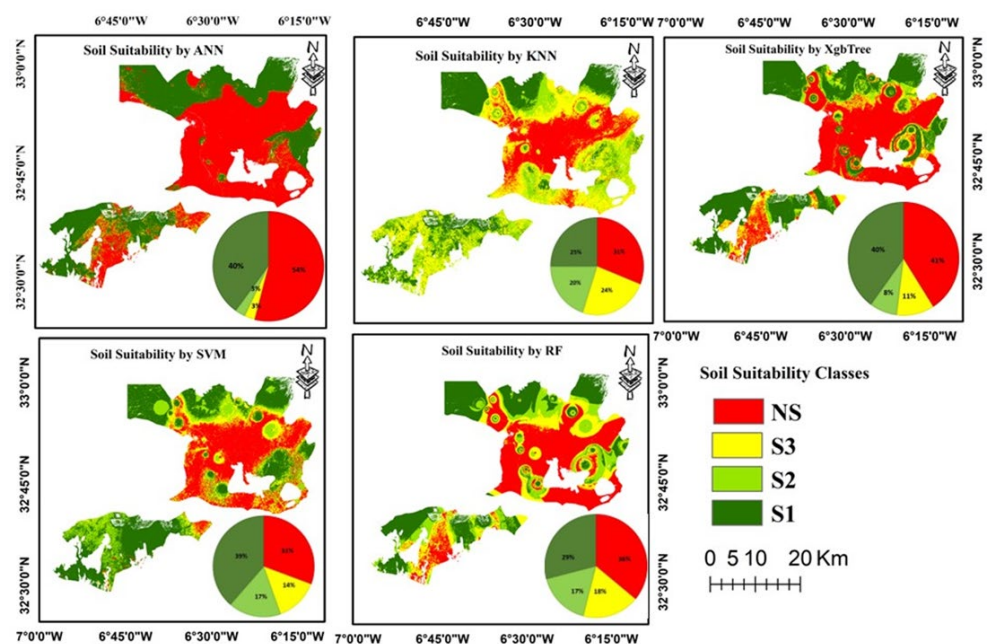


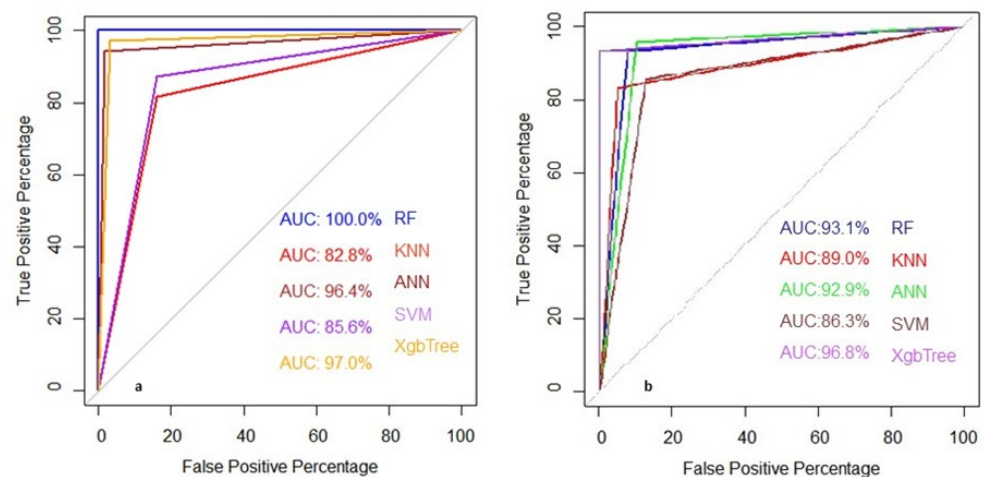
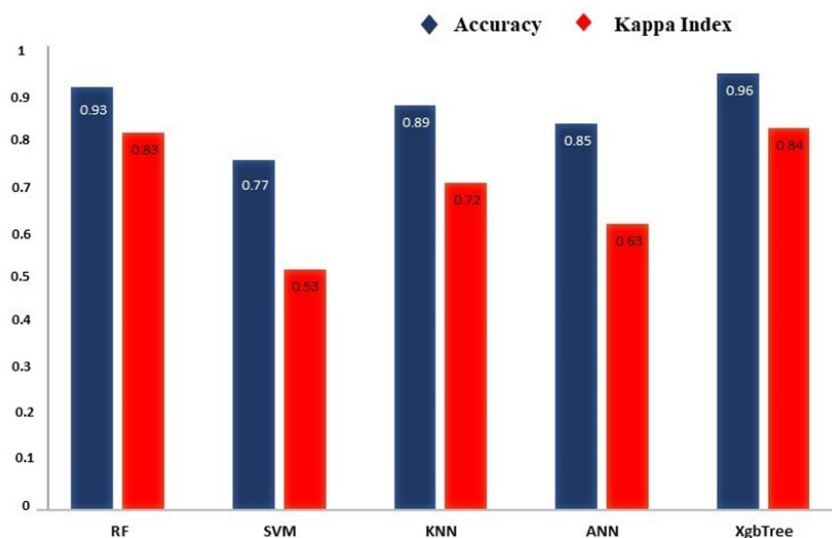
Figure 7. Soil Suitability maps using ML Models.

**Table 3.** The percentage of suitability classes in the five models.

Soil Suitability	ANN	KNN	RF	SVM	XgbTree
NS	54%	31%	36%	31%	41%
S3	3%	24%	18%	14%	11%
S2	3%	20%	17%	17%	8%
S1	40%	25%	29%	39%	40%

### 3.3. Validation of the Models

The validation of the results was carried out using ROC curves and AUC values of RF, KNN, ANN, XgbTree, and SVM models (Figures 8 and 9). We found a good similarity between the data collected during the fieldwork and the predicted results. The success rates and the predictive rates of the RF, KNN, ANN, SVM, and XgbTree models are 0.93, 0.89, 0.92, 0.86, and 0.96 and 1, 0.82, 0.96, 0.85, and 0.97, respectively. The AUCs indicate a very good to excellent prediction accuracy of the models for the SSM, and the XgbTree is the most accurate and robust model for mapping the soil suitability.

**Figure 8.** ROC curves showing AUC of RF, KNN, ANN, SVM, and XgbTree models (a) training dataset and (b) validation dataset.**Figure 9.** Bar plot for Accuracy and Kappa index.

#### 4. Discussion

Considerable progress has been made using the combined data coming from satellites and punctual observation processed with ML methods to map agricultural land potential [27]. Soil suitability assessment based on effective geo-environmental factors is the first step for managing soil suitability. Different approaches and procedures for the spatial prediction of the potential of soil have been developed and implemented around the world [9,31,41]. A controversial issue among environmental researchers is the preparation of a logical and reliable suitability map of soil. In the past decade, machine learning techniques were developed to make several important applications, such as the prediction, categorization, clustering, and elaboration of data [1]. Different sources were used to prepare the input dataset. As some of the factors considered in the SSM were derived from an NDVI time series, the resolution of the NDVI greatly affects the precision of the results [52]. In this study RF, ANN, KNN, SVM, and XgbTree-based machine learning algorithms were used to produce soil suitability maps based on training and validation datasets and 18 factors, including phenological and soil factors. The high values for the effective parameters demonstrate a significant correlation with soil suitability probability. As per the RF and KNN models, the most effective factors are CaCO<sub>3</sub>, K<sub>2</sub>O, ESP, Na, and phenological parameters, the SSMs based on the machine learning ensemble models, namely the RF, XgbTree, KNN, ANN, and SVM. According to field information, these areas represent very suitable soil with water resources considered to be the only limiting factor to overcome using irrigation [36,47,50]. The class S3 area represents 24.39% of the total area according to the model map with the NS in the models map area decreases by almost 20%. About less than 54% Table 3 of the total study area was classified unsuitable using the ANN approach when compared to 35.15% generated by the RF and xgbtree models Figures 7 and 9. This result can be explained by the dependency of this classification upon climate change, which can be caused by an increase in the extreme phenomena of precipitation and temperature that mainly affects the arid and semi-arid zones [36]. On the contrary, the ML approach aimed to meet all the climatic change needs by using more than 18 factors, in order to remove all obstacles and keep only the soil suitability. In Figure 7, the results of all the methods were explored and evaluated by comparing the suitability classes of the models with ground truth [42]. Through a visual comparison, we found that the result of the RF is close to reality as it shows that the soil is unsuitable (NS) for agriculture purposes in areas of phosphate mining activity within the study area; whereas, the ANN model shows, as a final result, that the soil is suitable for production in these areas. In the second part, we find a big difference in the result, where the ANN model shows that the soil is unsuitable for production, while the other models show a high capacity of the soil production also identifying some irrigated areas which already exist, which confirms the high suitability of these areas for intensive crop production.

#### 5. Conclusions

The purpose of this study is not only to investigate the capability of a machine learning model to predict the soil suitability, but also to compare its capability and robustness among the implemented models, i.e., XgbTree, RF, KNN, ANN, and SVM. Therefore, 16 geo-environmental and phenological factors were used and the significance of all the SSMs was explored using the previous models. The findings highlighted that understanding the strengths and limitations of each model remains difficult, even when performing model comparisons with specific goals in mind, such as prediction accuracy and robustness. Based on six threshold-dependent and independent assessment criteria, the RF and XgbTree models achieved the best results. The SVM, ANN, and KNN have a slightly lower precision when compared to the RF and XgbTree models in terms of pure prediction performance. The results of all the models show that the north part of the study area has the highest suitability. The outcome of the variable significance showed that the phenological parameters have the most significant soil suitability followed by the influences of chemical factors. On the other hand, the geology, pH, and soil salinity influences are the least important. The results

of this research can be helpful for land resource management to cope with the current uncertain situation and more accurately understand the different factors that influence soil suitability. Additionally, this approach can be used as a guideline for future research to analyze the capacity of soil in land use i.e., as a tool for regional soil resource analysis and a purely remote method in the way of developing.

**Author Contributions:** Conceptualization, M.I., S.K. and M.N.; Funding acquisition, K.A. and M.S.F.; Methodology, M.I. and M.N.; Project administration, T.B.; Resources, K.A., M.S.F. and T.B.; Software, M.I. and A.H.; Supervision, S.K., M.N. and T.B.; Validation, M.I. and M.N.; Visualization, M.I., M.N. and A.H.; Writing—original draft, M.I.; Writing—review & editing, M.I., S.K., M.N., A.H., R.L., H.E., E.F. and T.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by Researchers Supporting Project number (RSP2023R351), King Saud University, Riyadh, Saudi Arabia.

**Data Availability Statement:** The data that support the findings of this study are openly available in GEE. The other soil data that support the findings of this study are available on request from the author, Tarik BENABELOUHAB data are not publicly available.

**Acknowledgments:** Deep thanks and gratitude to the Researchers Supporting Project number (RSP2023R351), King Saud University, Riyadh, Saudi Arabia for funding this re-search article.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Abd-Elmabod, S.; Bakr, N.; Muñoz-Rojas, M.; Pereira, P.; Zhang, Z.; Cerdà, A.; Jordán, A.; Mansour, H.; De La Rosa, D.; Jones, L. Assessment of Soil Suitability for Improvement of Soil Factors and Agricultural Management. *Sustainability* **2019**, *11*, 1588. [CrossRef]
2. Hanh, H.Q.; Azadi, H.; Dogot, T.; Ton, V.D.; LeBailly, P. Dynamics of Agrarian Systems and Land Use Change in North Vietnam. *Land Degrad. Dev.* **2017**, *28*, 799–810. [CrossRef]
3. Santana-Cordero, A.M.; Ariza, E.; Romagosa, F. Studying the historical evolution of ecosystem services to inform management policies for developed shorelines. *Environ. Sci. Policy* **2016**, *64*, 18–29. [CrossRef]
4. Tengberg, A.; Radstake, F.; Zhang, K.; Dunn, B.; Tengberg, A. Scaling up of Sustainable Land Management in the Western People's Republic of China: Evaluation of a 10-Year Partnership. *Land Degrad. Dev.* **2013**, *27*, 134–144. [CrossRef]
5. Allbed, A.; Kumar, L. Soil Salinity Mapping and Monitoring in Arid and Semi-Arid Regions Using Remote Sensing Technology: A Review. *Adv. Remote Sens.* **2013**, *2*, 373–385. [CrossRef]
6. Brevik, E.C.; Calzolari, C.; Miller, B.A.; Pereira, P.; Kabala, C.; Baumgarten, A.; Jordán, A. Soil mapping, classification, and pedologic modeling: History and future directions. *Geoderma* **2016**, *264*, 256–274. [CrossRef]
7. Panagos, P.; Borrelli, P.; Poesen, J.; Ballabio, C.; Lugato, E.; Meusburger, K.; Montanarella, L.; Alewell, C. The new assessment of soil loss by water erosion in Europe. *Environ. Sci. Policy* **2015**, *54*, 438–447. [CrossRef]
8. Sam, K.; Coulon, F.; Prpich, G. Working towards an integrated land contamination management framework for Nigeria. *Sci. Total Environ.* **2016**, *571*, 916–925. [CrossRef]
9. Velmurugan, A.; Swarnam, T.; Ambast, S.; Kumar, N. Managing waterlogging and soil salinity with a permanent raised bed and furrow system in coastal lowlands of humid tropics. *Agric. Water Manag.* **2016**, *168*, 56–67. [CrossRef]
10. Adeyolanu, O.D.; Are, K.S.; Adelana, A.O.; Denton, O.A.; Oluwatosin, G.A. Characterization, Suitability Evaluation and Soil Quality Assessment of Three Soils of Sedimentary Formation for Sustainable Crop Production. *J. Agric. Ecol. Res. Int.* **2017**, *11*, 1–10. [CrossRef]
11. Rossiter, D. Land evaluation: Towards a revised framework; Land and Water Discussion Paper 6, FAO. FAO, Rome (2007), 107 pp., ISSN: 1729-0554; Only available in PDF format as [www.fao.org/nr/lman/docs/lman\\_070601\\_en.pdf](http://www.fao.org/nr/lman/docs/lman_070601_en.pdf); free. *Geoderma* **2009**, *148*, 428–429. [CrossRef]
12. Pereira, P.; Brevik, E.; Trevisani, S. Mapping the environment. *Sci. Total Environ.* **2018**, *610–611*, 17–23. [CrossRef]
13. Smetanova, A.; Pereira, P.; Brevik, E.; Munoz-Rojas, M.; Miller, B.; Smetanova, A.; Depellegrin, D.; Misiune, I.; Novara, A.; Cerda, A. Soil mapping and process modelling for sustainable land management. In *Soil Mapping and Process Modelling for Sustainable Land Use Management*; Pereira, P., Brevik, E., Munoz-Rojas, M., Miller, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2017; pp. 29–60, ISBN 9780128052006.
14. Ghosh, P.; Lepcha, K. Weighted linear combination method versus grid based overlay operation method—A study for potential soil erosion susceptibility analysis of Malda district (West Bengal) in India. *Egypt. J. Remote Sens. Space Sci.* **2018**, *22*, 95–115. [CrossRef]
15. Oertli, J.J. Limitations to the diagnostic information obtained from soil analyses. *Nutr. Cycl. Agroecosystems* **1990**, *26*, 189–196. [CrossRef]

16. Akumu, C.; Johnson, J.; Etheridge, D.; Uhlig, P.; Woods, M.; Pitt, D.; McMurray, S. GIS-fuzzy logic-based approach in modeling soil texture: Using parts of the Clay Belt and Hornepayne region in Ontario Canada as a case study. *Geoderma* **2015**, *239–240*, 13–24. [CrossRef]
17. Habibi, V.; Ahmadi, H.; Jafari, M.; Moeini, A. Mapping soil salinity using a combined spectral and topographical index with artificial neural network. *PLoS ONE* **2021**, *16*, e0228494. [CrossRef]
18. Whitney, K.; Scudiero, E.; El-Askary, H.M.; Skaggs, T.H.; Allali, M.; Corwin, D.L. Validating the use of MODIS time series for salinity assessment over agricultural soils in California, USA. *Ecol. Indic.* **2018**, *93*, 889–898. [CrossRef]
19. Zhang, T.-T.; Qi, J.-G.; Gao, Y.; Ouyang, Z.-T.; Zeng, S.-L.; Zhao, B. Detecting soil salinity with MODIS time series VI data. *Ecol. Indic.* **2015**, *52*, 480–489. [CrossRef]
20. Ali, I.; Greifeneder, F.; Stamenkovic, J.; Neumann, M.; Notarnicola, C. Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data. *Remote Sens.* **2015**, *7*, 16398–16421. [CrossRef]
21. Malczewski, J. Ordered weighted averaging with fuzzy quantifiers: GIS-based multicriteria evaluation for land-use suitability analysis. *Int. J. Appl. Earth Obs. Geoinf.* **2006**, *8*, 270–277. [CrossRef]
22. Mokarram, M.; Hojati, M. Using ordered weight averaging (OWA) aggregation for multi-criteria soil fertility evaluation by GIS (case study: Southeast Iran). *Comput. Electron. Agric.* **2017**, *132*, 1–13. [CrossRef]
23. Barakat, A.; Hilali, A.; El Baghdadi, M.; Touhami, F. Landfill site selection with GIS-based multi-criteria evaluation technique. A case study in Béni Mellal-Khouribga Region, Morocco. *Environ. Earth Sci.* **2017**, *76*, 413. [CrossRef]
24. McBratney, A.; Santos, M.M.; Minasny, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
25. Roell, Y.E.; Beucher, A.; Möller, P.G.; Greve, M.B.; Greve, M.H. Comparing a Random Forest Based Prediction of Winter Wheat Yield to Historical Yield Potential. *Agronomy* **2020**, *10*, 395. [CrossRef]
26. Taghizadeh-Mehrjardi, R.; Nabiollahi, K.; Rasoli, L.; Kerry, R.; Scholten, T. Land Suitability Assessment and Agricultural Production Sustainability Using Machine Learning Models. Available online: <https://doaj.org/article/84f719cf0a2549d0a85bdce76ed12a97> (accessed on 3 October 2022).
27. Rentschler, T.; Gries, P.; Behrens, T.; Bruelheide, H.; Kühn, P.; Seitz, S.; Shi, X.; Trogisch, S.; Scholten, T.; Schmidt, K. Comparison of catchment scale 3D and 2.5D modelling of soil organic carbon stocks in Jiangxi Province, PR China. *PLoS ONE* **2019**, *14*, e0220881. [CrossRef]
28. Teng, H.; Rossel, R.A.V.; Shi, Z.; Behrens, T. Updating a national soil classification with spectroscopic predictions and digital soil mapping. *Catena* **2018**, *164*, 125–134. [CrossRef]
29. West, D.; Dellana, S.; Qian, J. Neural network ensemble strategies for financial decision applications. *Comput. Oper. Res.* **2005**, *32*, 2543–2559. [CrossRef]
30. Jayaraman, V.; Sridevi, S.; Monica, K.M.; Lakshminarayanan, A.R. Predicting the Soil Suitability using Machine Learning Techniques. In Proceedings of the 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, 19–21 November 2021; Volume 1, pp. 200–202. [CrossRef]
31. Chan, J.C.-W.; Paelinckx, D. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sens. Environ.* **2008**, *112*, 2999–3011. [CrossRef]
32. Wang, S.-J.; Mathew, A.; Chen, Y.; Xi, L.-F.; Ma, L.; Lee, J. Empirical analysis of support vector machine ensemble classifiers. *Expert Syst. Appl.* **2009**, *36*, 6466–6476. [CrossRef]
33. Pham, B.T.; Nguyen, M.D.; Nguyen-Thoi, T.; Ho, L.S.; Koopialipoor, M.; Quoc, N.K.; Armaghani, D.J.; Van Le, H. A novel approach for classification of soils based on laboratory tests using Adaboost, Tree and ANN modeling. *Transp. Geotech.* **2020**, *27*, 100508. [CrossRef]
34. Sheikhi, S. An effective fake news detection method using WOA-xgbTree algorithm and content-based features. *Appl. Soft Comput.* **2021**, *109*, 107559. [CrossRef]
35. Benabdelouahab, T.; Lebrini, Y.; Boudhar, A.; Hadria, R.; Htitiou, A.; Lionbou, H. Monitoring spatial variability and trends of wheat grain yield over the main cereal regions in Morocco: A remote-based tool for planning and adjusting policies. *Geocarto Int.* **2019**, *36*, 2303–2322. [CrossRef]
36. Diouf, A.A.; Hiernaux, P.; Brandt, M.; Faye, G.; Djaby, B.; Diop, M.B.; Ndione, J.A.; Tychon, B. Do Agrometeorological Data Improve Optical Satellite-Based Estimations of the Herbaceous Yield in Sahelian Semi-Arid Ecosystems? *Remote Sens.* **2016**, *8*, 668. [CrossRef]
37. Barakate, M.; Ouhdouch, Y.; Oufdou, K.; Beaulieu, C. Characterization of rhizospheric soil streptomycetes from Moroccan habitats and their antimicrobial activities. *World J. Microbiol. Biotechnol.* **2002**, *18*, 49–54. [CrossRef]
38. Khellouk, R.; Barakat, A.; El Jazouli, A.; Boudhar, A.; Lionbou, H.; Rais, J.; Benabdelouahab, T. An integrated methodology for surface soil moisture estimating using remote sensing data approach. *Geocarto Int.* **2019**, *36*, 1443–1458. [CrossRef]
39. Rapport Gebral, Direction Regional de L'agriculture Monographie de la Region Beni Mellal Khenifra. Marchè N 25/99-00/DPA/38/SA 2015. Beni Mellal, Morocco, 2015. Available online: <https://coeurdumaroc.ma/cri/public/documents/agriculture-72745.pdf> (accessed on 29 November 2022).
40. Namous, M.; Hssaisoune, M.; Pradhan, B.; Lee, C.-W.; Alamri, A.; Elaloui, A.; Edahbi, M.; Krimissa, S.; Eloudi, H.; Ouayah, M.; et al. Spatial Prediction of Groundwater Potentiality in Large Semi-Arid and Karstic Mountainous Region Using Machine Learning Models. *Water* **2021**, *13*, 2273. [CrossRef]

41. Pulatov, A.; Khamidov, A.; Akhmatov, D.; Pulatov, B.; Vasenev, V. Soil salinity mapping by different interpolation methods in Mirzaabad district, Syrdarya Province. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *883*, 012089. [[CrossRef](#)]
42. Jonsson, P.; Eklundh, L. Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 1824–1832. [[CrossRef](#)]
43. Reed, B.C.; Brown, J.F.; Vanderzee, D.; Loveland, T.R.; Merchant, J.W.; Ohlen, D.O. Measuring phenological variability from satellite imagery. *J. Veg. Sci.* **1994**, *5*, 703–714. [[CrossRef](#)]
44. Durgun, Y.; Gobin, A.; Van De Kerchove, R.; Tychon, B. Crop Area Mapping Using 100-m Proba-V Time Series. *Remote Sens.* **2016**, *8*, 585. [[CrossRef](#)]
45. Htitiou, A.; Boudhar, A.; Chehbouni, A.; Benabdelouahab, T. National-Scale Cropland Mapping Based on Phenological Metrics, Environmental Covariates, and Machine Learning on Google Earth Engine. *Remote Sens.* **2021**, *13*, 4378. [[CrossRef](#)]
46. Htitiou, A.; Boudhar, A.; Lebrini, Y.; Hadria, R.; Lionboui, H.; Elmansouri, L.; Tychon, B.; Benabdelouahab, T. The Performance of Random Forest Classification Based on Phenological Metrics Derived from Sentinel-2 and Landsat 8 to Map Crop Cover in an Irrigated Semi-arid Region. *Remote Sens. Earth Syst. Sci.* **2019**, *2*, 208–224. [[CrossRef](#)]
47. Diouf, A.A.; Djaby, B.; Diop, M.B.; Wele, A.; Ndione, J.A. Tychon Fonctions D’ajustement Pour L’estimation de la Production Fourragère Herbacée des Parcours Naturels du Sénégal à Partir du NDVI s10 de SPOT-Vegetation. Juill. 2014. Available online: <https://orbi.uliege.be/handle/2268/203858> (accessed on 21 May 2019).
48. Lebrini, Y.; Boudhar, A.; Hadria, R.; Lionboui, H.; Elmansouri, L.; Arrach, R.; Ceccato, P.; Benabdelouahab, T. Identifying Agricultural Systems Using SVM Classification Approach Based on Phenological Metrics in a Semi-arid Region of Morocco. *Earth Syst. Environ.* **2019**, *3*, 277–288. [[CrossRef](#)]
49. Brownlee, J. Boosting and AdaBoost for Machine Learning. *Machine Learning Mastery*. 24 April 2016. Available online: <https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/> (accessed on 22 August 2022).
50. Kawabata, D.; Bandibas, J. Landslide susceptibility mapping using geological data, a DEM from ASTER images and an Artificial Neural Network (ANN). *Geomorphology* **2009**, *113*, 97–109. [[CrossRef](#)]
51. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
52. Yang, L.; He, X.; Shen, F.; Zhou, C.; Zhu, A.-X.; Gao, B.; Chen, Z.; Li, M. Improving prediction of soil organic carbon content in croplands using phenological parameters extracted from NDVI time series data. *Soil Tillage Res.* **2019**, *196*, 104465. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.