AI4EO hyperview challenge: combination of machine learning methods on hyperspectral images to predict the soil parameters

Marsia Sanità[*], Eva Savina Malinverni[*], Roberto Pierdicca[*], Adriano Mancini[**], Ewa Glowienka[***], Lindo Nepi[**]

[*] Dipartimento di Ingegneria Civile, Edile e dell’Architettura (DICEA), Università Politecnica delle Marche, 60100 Ancona, Italy; m.sanita@pm.univpm.it, e.s.malinverni@staff.univpm.it, r.pierdicca@staff.univpm.it,
[**]Department of Information Engineering (DII), Università Politecnica delle Marche, 60131, Ancona, Italy; l.nepi@pm.univpm.it, a.mancini@univpm.it,
[***]Faculty of Mining Surveying and Environmental Engineering, Department of Photogrammetry, Remote Sensing of Environment, and Spatial Engineering, AGH University of Science and Technology, al. Mickie wicza 30, 30-059 Krakow, Poland; eglo@agh.edu.pl

Keywords: Remote Sensing, Machine Learning Multi-output Regression Model, Hyperspectral, Soil Parameters.

Abstract

In the AI4EO educational challenge "Seeing Beyond the Visible", hyperspectral images are used to predict the chemical parameters on the soil (K, Mg, P_2O_5, pH) in anticipation of the correct use of fertilisers. The challenge is set in an agricultural area of Poland and the available data are hyperspectral images (150 contiguous hyperspectral bands) and in situ samples for soil parameter measurements. The aim of this challenge was to advance the state of art of soil parameter analysis by hyperspectral images. Having a good knowledge of the chemical characteristics of the soil is important in order to be able to identify which types of crops are most suitable in that area to optimise production and reduce the use of fertilisers. In the face of ongoing climate change and the disastrous calamitous events that follow, the idea of a sustainable agriculture becomes a necessity. Artificial intelligence (AI) through Machine Learning (ML) and Deep Learning (DL) techniques can be a great support for farmers in optimising the use of natural resources and ensuring better land management. In this paper, a group of engineers in the field of data science and geomatics carries out this research topic accepting the challenge proposed by AI4EO. A variety of AI techniques were applied by the authors of this paper with respect to the other participants in the challenge methods. The proposed approach is based on the novelty of a dataset filtering and on the use of a Random Forest Multi-Output Regressor.

1. Introduction

Due to the ongoing climate emergency, the balance of the world's ecosystem is so precarious that it is no longer certain guarantee food for the entire world population. For this reason, sustainable and precision agriculture (PA) is being promoted, which pays more attention to the chemical characteristics of the soil and the type of cultivation it is suited to to avoid the use of pesticides and fertilizers as much as possible. Climate change not only brings about a change in temperatures, it also causes a change in the percentage of nutrients in the soil. The phenomenon that is worrying is the parallelism that is taking place: on the one hand a fast population growth (Ghosh et al., 2024; Dhiman et al., 2023) and on the other a decrease in agricultural production of cereals such as maize (Ocwa et al., 2023). In September 2015 the Agenda 2030 was signed and it consists of 17 Sustainable Development Goals (SDGs) and each of them has a related challenge (ASVIS). The Goal 2 takes up that cause being involved in shortage and malnutrition. Among the targets listed under Goal 2 are the doubling of agricultural production, the implementation of sustainable agriculture and the use of resilient agricultural practices. The first shrewdness to be taken is certainly try to limit the use of pesticides, herbicides and insecticides and produce as naturally as possible while respecting biodiversity as much as possible (Pandey et al., 2023). The coupling of technological usages, remote sensing and agricultural practices are a possibility to try to mediate these issues. An opportunity to be considered is the potential of satellite remote sensing or drone imagery. The possibility of combining free satellite images to make a prediction of the amount and type of nutrients in the soil exists is a trend in the literature.

In this work it is exploited the potential of hyperspectral images capable of capturing images with multiple optical bands useful to predict soil parameters which is the aim of the challenge “‘Seeing Beyond the Visible” (The European Space Agency; Nalepa et al., 2022; Nalepa et al., 2024). Hyperspectral remote sensing is therefore very useful for precision farming and planning. The challenge provides a dataset consisting of hyperspectral imagery and ground sampling data (Section 2.2). The solution proposed in this paper is based on the use of ML algorithms and Neural Networks (NN). The best solution is achieved by Random Forest Multi-Output Regressor model performance.

The paper is structured as it follows: in the first section, there is the introduction and the state of art of that topic. Materials and methods compose the second section in which there is a short description of the AI4EO challenge and the available dataset, followed by the methodology, the model and the innovative approach proposed by the authors’ team. The third section is the results and discussion, and the last one is the conclusion.

1.1 State of the art

The idea of considering non-invasive and inexpensive solutions would help broaden the scope of the analysis to be performed

given the spatial resolution that a satellite image can offer. The importance of being able to estimate the amount of nutrients in the soil is also justified by their actual chemical imbalance. The current condition is that of extremely fertilized areas, cause of environmental pollution, and areas with low production due to the poverty of nutrients (Penuelas et al., 2023). It must be taken into consideration that soil plays a fundamental function within the complex climate system. Due to global warming, increased rainfall and catastrophic events, the yield of agricultural fields has decreased and so has the productivity of livestock (Moersdorf et al., 2024; Santosh et al., 2023; Balasundram et al., 2023; Bibi et al., 2023). From the point of view of environmental sustainability, it must be noted that almost a quarter of greenhouse gas emissions have agriculture as their source (Filho et al., 2023). It will ensure a fair balance between respecting natural resources such as water, land and biodiversity and ensuring the necessary nourishment for living beings. In the last 30 years, the idea of digital agriculture (DA) has developed with the aim of increasing production efficiency (Balasundram et al., 2023). It includes various tools and technologies as IoT systems or Apps for continuous monitoring of water or land conditions (Sarfraz et al., 2023). As said before it is possible to obtain thematic maps useful for identifying critical or stressed areas by monitoring thought satellite or drone images (Balasundram et al., 2023). An example of how to use aerial photos and photos interpretations to analyze the rural landscapes is conducted in (Chiappini et al., 2023). A very strong contribution, although it can be costly, as said before, is made by hyperspectral imagery, whose role is strictly useful in achieving PA providing useful information to improve cultivation practices. In (Pande et al., 2023) is an example of coupling information from hyperspectral data with geospatial data on landforms using (slope, soil nutrient, crop quality etc.,) in a Geo-graphical Information System (GIS) environment. The biggest challenge is to be able to exploit satellite data for the estimation of the soil parameters. The difficulty lies in the non-linearity of the problem; so, to overcome this problem, some researchers applied machine learning algorithms to predict the component. In (Yu et al., 2023; Mukhamediev et al., 2023; Song et al., 2023; Zayani et al., 2023) several machine learning (ML) algorithms were used for the prediction of the Soil Organic Matter (SOM) and salinity. The quantity of (SOM) plays an important role in terms of productivity. To control soil properties the main applications involve the collection of the sample manually and then its analysis in the laboratory. This procedure requires a considerable expense, so remote sensing techniques can be a viable alternative in detecting the quantity of SOM. The soil has its own spectral profile based on the combination of its constituents and their quantity. Since the agricultural yield strictly depends on the presence of some parameters in the soil such as nitrogen (N), phosphorus (P), potassium (K), pH, soil temperature and its humidity, the need for continuous monitoring of them is inevitable. As mentioned above, one possibility is the use of IoT systems that is based on ML techniques. In this way the system will be able to recommend the best type of crop to apply in that specific agricultural space, reducing the quantity of fertilizers to a minimum (Hossain et al., 2023).

2. Materials and Methods

2.1 AI4EO Challenge "Seeing Beyond the Visible"

"Seeing Beyond the Visible" is a permanent educational challenge and involves hyperspectral data (The European Space Agency; Nalepa et al., 2022; Nalepa et al., 2024). The challenge is based on the achievement of a score by means of the use of models applied by each team in comparison with the reference value of baseline Regressor model. The success of the challenge depends on the lower values achieved by the choice of team model. It is available on ESA’s AI4EO platform, and it was launched by KP Labs together with ESA (European Space Agency) and partner QZ Solutions (The European Space Agency; Nalepa et al., 2022; Nalepa et al., 2024). It is an extraordinary challenge that gives the possibility of estimating soil parameters from hyperspectral data. It can significantly revolutionize the world of agriculture. The support guarantee to help farmers in the correct management of the soil, the distribution of fertilizers or even lead to the elimination of their use. The objective of this challenge is to advance the state of the art on this topic on the analysis of hyperspectral images for the soil parameters prediction.

2.2 Dataset

The aim of the challenge is in anticipation of the launch of the Intuition-1 mission (The European Space Agency; Nalepa et al., 2022; Nalepa et al., 2024). The dataset includes data from on-site samples and data from hyperspectral measurements collected in flight. The in-situ measurements of the soil parameters refer to a period from August 2020 to November 2020 and are used as ground-truth. The method used for the extraction of these samples is Mehlich 3 and it is presented in (The European Space Agency; Nalepa et al., 2022; Nalepa et al., 2024). For each sample area, covering from 0.5 to 4 hectares, 12 samples have been collected. Once the sample has been extracted, it has been taken to the laboratory and from each one of them the value of the 4 reference parameters (pH, K, P_2O_5 and Mg) has been obtained. The hyperspectral sensor has been mounted on the Piper PA-31 Navajo aircraft (with the 2550–2700 m flying altitude, 61.8 m/s cruising speed, 2 m GSD, cloudless and windless weather) (The European Space Agency; Nalepa et al., 2022; Nalepa et al., 2024). The acquisition campaign was conducted in March 2021 on an agricultural area of Poland. To capture the hyperspectral images the system used the HySpex VS-725 (Norsk Elektro Optikk AS) composed by two SWIR-384 sensors and ones VNIR-1800 sensor able to work simultaneously (Table 1).

The hyperspectral data contains 150 contiguous bands (462-942 nm, with a spectral resolution of 3.2 nm), which reflects the spectral range of the hyperspectral imaging sensor deployed on-board Intuition-1. The Intuition-1 satellite mission allows to

	SWIR-384	VNIR-1800
Spectral range	930-2500 nm	400-1000 nm
Number of bands	288	186
Spectral resolution	5.45 nm	3.26 nm

Table 1. Sensors characteristics

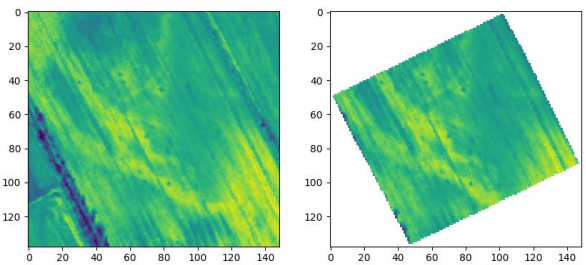

Figure 1. Representation of band 120 (842.47 nm).

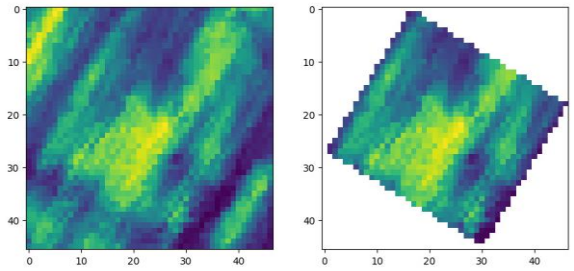

Figure 2. Representation of band 100 (778.54 nm).

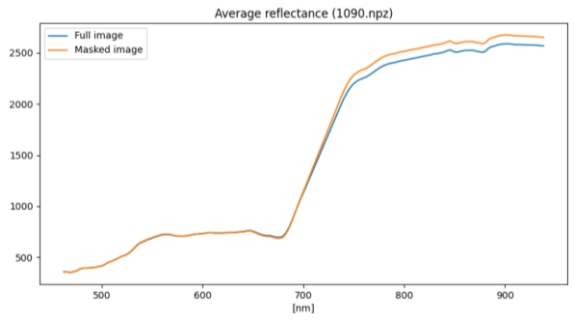

Figure 3. Representation of average reflectance (1090.npz).

observe the earth through a hyperspectral instrument and through an on-board computing system capable of processing the data with a neural network directly in orbit (KP LABS). Intuition-1 is a demonstrator of how AI positively influences the RS process. In addition, the use of a hyperspectral instrument allows the detection of phenomena that panchromatic, multispectral and RADAR instruments are not able to do so far. The dataset available for this challenge is composed of a total of 1732 patches used for training and a total of 1154 patches for testing. Each patch has a size on average around 60 x 60 pixels; it depends on agricultural parcels. An example of the data available in the challenge are the Fig. 1 and Fig. 2 in which are represented respectively the band 120 and band 100 and Fig. 3 that represents the average reflectance of the sample 1090. Fig. 2 and Fig. 3 represent both two images: the left is the full version of the image, while the right is the masked version. The masked array are all the array that may have missing or invalid entries. The training dataset, without any skimming (Section 2.5) contains 1500 samples, while the validation dataset (used to determine the Score) contains 232 samples. The sample presented in Fig. 3 represents a training sample.

2.3 Methodology

Machine learning algorithms make it possible to identify relationships between data without using explicit instructions by exploiting a training dataset. In the case of the challenge of that paper, it is the condition of supervised machine learning technique. Supervised learning algorithms are trained on labelled datasets. For each input dataset into the algorithm, the corresponding output dataset is provided. This allows to solve classification or value estimation (Regression) tasks comparing various algorithms, such as Decision Trees (DT), Linear Regression (LR), K-nearest neighbors (K-NN), Random Forest (RF), XGBoost etc.,) until a sufficiently accurate solution is obtained. In Fig. 4 is represented the methodology applied. In scheme a) is represented the general approach instead in b) is represented the author’s approach.

In this case, having multiple target variables simultaneously, the model has to be based on multi-target regressor scheme (Fig. 5). Their multi-target functionality is supported by several supervised ML algorithms as RF, DT, L R and K-NN.

It is necessary to make a prediction of the values of the four soil parameters that are the dependent variables (y1, y2, y3 and y4) coming from the analysis of the bands (x1,…, xn). To perform this task, has been decided to use different ML models implemented in the scikit-learn Python package, using default values for different parameters. The algorithms used are: DT, LR, K-NN and RF. In addition to these algorithms, some 'fully connected' neural networks have been created, with different numbers of layers. Neural networks are used in many fields, and in recent years it has been seen their power applied to solving difficult tasks such as computer vision, speech recognition and text generation. In this work, after using various ML algorithms, it has been decided to use neural networks to process the input data of the challenge. Neural networks are used extensively in difficult tasks due to their ability to automatically extract complex features from the data, without a manually “feature extraction” phase as in ML algorithms. Neural networks can therefore achieve greater accuracy than traditional ML models, however, the results depend on several factors, such as the quality and quantity of the data and the architecture of the neural network. In this case, the authors have modified the structure of the neural network by varying the number of hidden layers, the values of the main hyperparameters (learning rate, weight decay, batch size) and using the dropout value. The technique of Batch Normalization (Batch Norm) is widely adopted in the case of neural networks to make the training phase faster and more stable. This is achieved by stabilizing the distributions of the input layers and introducing additional layers that control mean and variance of these distributions. The best-performing values determined experimentally have been

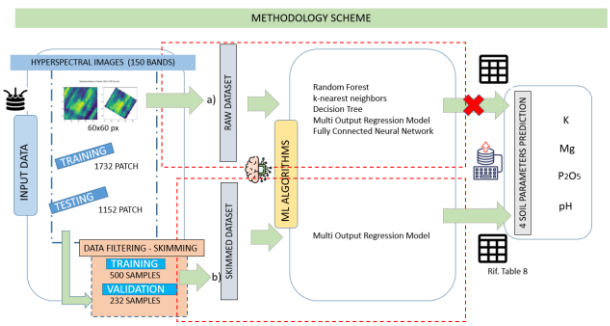

Figure 4. Methodology scheme a) raw dataset; b) authors’ approach based on dataset filtering.

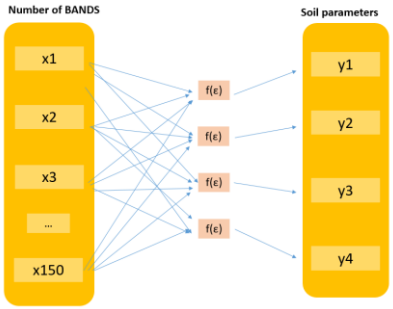

Figure 5. Multi-Output Regression model scheme.

learning rate=0.00006, weight decay=0.001, activation function=Tanh, dropout=0.4. The learning rate value has a fundamental role in the model because from it depends on the regulations of the parameters. Its value is in the range between 0 and 1. The choice of the best learning rate value is achieved through a trade-off between convergence speed and optimizer stability. In the case of the high convergence rate, the risk is to have an algorithm that is not stable. In the case of slow convergence, on the other hand, the risk is an overlay stable situation. The weight decay comes from 0 to 0.1 on a logarithmic scale and it is used to improve the model performance. Its value depends on various factors such as the complexity of the model and the amount of the training data. Changing the number of hidden layers leads to considerable variations in performance; in some cases, worst results are obtained respect to the baseline model. In neural networks, the number of hidden layers to be used in each task is not defined a priori.

2.4 Model

Experimentally, the best number of layers of a neural network depends on the complexity of the problem, although it does not guarantee greater accuracy of the result. A greater number of hidden layers, appropriately connected, allows the non-linear relationships between the input features (values relative to the 150 spectral bands) and the output of the problem (4 values of the soil parameters) to be learned with greater accuracy. It has been proceeded by constructing a network with a small number of hidden layers (1 or 2 layers), increased, if necessary, to constantly monitor the performance and assess whether overfitting or underfitting. In this model, the best performance has been obtained with a structure illustrated in Table 2.

In this model, a Batch Norm layer has been used, which greatly improved the accuracy of the result (Santurkar et al., 2018).

In relation to the number of neurons (60), in the layers, the best performance has been obtained with approximately half the number of them in input feature (150). The results obtained from the various models analysed are shown in Table 3. The scores values achieved by using the Multi-Output Regressor model without the selection of the input data are showed in Table 3. The last line of the Table 3 shows the values achieved by using the neural network considered in this paper. The results obtained with this neural network with Batch Norm, compared to those achieved with the other algorithms have a better score, even reaching a value of 0.99 for P_2O_5. However, it should be considered that these values are the result of a neural network that has worked with all input data without previously skimming them (Section 2.5). The input data are composed by a training dataset (1500 samples) and a testing dataset (232 samples). The limit is that the performance remained unchanged and very high with any input data. The difference in the method in this paper respect to the others is in the skimming of the input data. The best hypothesis has been to choose to skim the input

Layer type	Output shape	Param #
dense_8 (Dense)	(None, 60)	9060
batch_normalization_3 (BatchNormalization)	(None, 60)	240
dropout_4 (Dropout)	(None, 60)	0
dense_9 (Dense)	(None,4)	244

Table 2. Neural Network model

Model	Class P_2O_5 score	Class K score	Class Mg score	Class pH score	Final score
Random Forest	1.4865	1.7038	1.2042	1.0607	1.3638
K-NN	1.6154	1.6248	1.5925	1.2271	1.5150
Linear Regression	1.3767	1.6441	1.1802	0.8096	1.2527
Decision Tree	2.5868	2.7199	2.3170	2.1736	2.4493
Multi-Output Regression model	1.4069	1.6025	1.0245	0.9493	1.2458
Fully Connected neural network	0.9953	0.4420	0.8115	0.7548	0.7509
Random Forest	1.4865	1.7038	1.2042	1.0607	1.3638

Table 3. Different applied AI model score.

dataset according to the low value of standard deviation. The baseline MSE reflects the performance of the algorithm that returns the average value of each soil parameter obtained for the training set. MSEi (Mean Squared Error) it is calculated according to the general MSE formula (Eq. 1) in which i refers to each soil parameter (i = 4) and it is used to calculate the score value (Eq. 2) of the model reported in Table 3 (The European Space Agency). The ψ parameter represents the cardinality of the test set.

$$MSEi = \frac{\sum_{j=1}^{|\Psi|}(p_j - \hat{p}_j)2}{|\Psi|} \quad (1)$$

$$Score = \frac{\sum_{i=1}^{4}(MSE_i / MSE_i^{base})}{4} \quad (2)$$

2.5 Approach to increase the accuracy of the model based on dataset filtering

Major headings Analysing the dataset, for each hyperspectral band, an average sample value recorded over the patch surveyed is extracted. The idea of considering a single average value for a patch that is not so small, can negatively affect the input dataset making noisy. Calculating the standard deviation for the various samples in the dataset provided by the authors of challenge, it has been seen that these values ranges from 16.8 to 705.6. For each sample, an average standard deviation value has been associated. Once, all the average values have been ordered in ascending order, our approach, based on data filtering, has been applied selecting the top 500 values. For this reason, the authors tried a new approach to perform the model in a more correct way to reduce that gap. The methodology undertaken includes to do a test to determine whether it is possible the reduction of the noise of the dataset, increasing the accuracy of the model. For all the samples, it is calculated the value of the data variance of each array representing each band. The samples in the dataset are then sorted in ascending order according to the average variances of the 150 hyperspectral bands, and then the 500 samples with the lowest average variances. From this subset of samples, a new array containing the first 200 samples is extracted. The dataset is then divided into two parts: training dataset and validation dataset.

The data provided used masked 2D arrays (numpy masked array) so the numpy.ma.std and numpy.ma.median from the

Numpy library are used to calculate the standard deviation and median. The standard deviation is used to select input data for a skimming dataset considering only samples with lowest dispersion while the median is then used on the skimmed samples in place of the average on each patch. Neural network models and other ML algorithms are then run on the new dataset consisting of extracted samples. The aim of approach is to show that the new dataset (skimmed dataset), which has less variance value, yields better results. So, the neural network used for the regression-based model result in an overfitting situation because the number of rows (samples) is too small compared to the features number (hyperspectral bands). It is therefore decided to use some ML models such as K-NN, DT and RF. A polynomial kernel PCA process is applied to the input dataset. The best performance is obtained by using the RF algorithm (parameter: n_estimators=100, oob_score=True, random_state=0), with Kernel PCA (polynomial kernel) after choosing the first 70 samples in order of increasing standard deviation for the training dataset.

3. Results and discussion

The approach proposed in this paper is based on the skimming of the input dataset considering only the values of the data with the smallest standard deviation. Also, in (KP LABS; Kuzu et al., 2022; Zelenka et al., 2022) the authors accepted the challenge and proposed their methodology. In regression problems, the presence of good MSE values and very low R^2 values is not an error but an indicator of a particular situation that requires further investigation, as has been done by the authors of this paper. MSE indicates how far away the model's predictions are from the actual values, but penalizes the largest deviations being MSE a quadratic error. A lower MSE value means that, on average, the model predictions are closer to the actual values. If the data have a particular distribution or wide variability, or if the model is excessively simple compared to the data, it is possible that the model makes more accurate predictions for some points causing a low MSE values. The authors’ analysis used ML models and NNs with different configurations, first on the original dataset and then on the variance-skimmed dataset, highlighted this data issue, which is therefore caused by the large variability of the original dataset. In Table 4 there are the results obtained using 70 samples with minor standard deviation. The average of the Score for each soil parameter is 0.96. Using a training dataset containing only 120 samples of the selected dataset in order of increasing standard deviation, the performance obtained is the worse (Table 5).

	P_2O_5	K	Mg	pH
MSE	563.711	3278.186	1267.657	645.429 e-02
MSE baseline	534.818	3164.276	1687.958	642.857 e-02
Score	1.054	1.036	0.751	1.004

Table 4. Results of 70 samples with the lower standard deviation

	P_2O_5	K	Mg	pH
MSE	875.108	2498.483	687.929	358.177e-02
MSE baseline	895.709	2581.077	760.142	399.826e-02
Score	0.9778	0.968	0.905	1.054

Table 5. Results of 120 samples with the lower standard deviation

To verify that the metrics are indeed influenced by the input dataset, an experiment has been conducted using 70 random samples without taking to account the standard deviation of the data. The results achieved are represented in Table 6. The algorithm used is a RF type with the same parameters as in the previous test and a Kernel PCA (kernel polynomial) is applied before of the implementation of the model.

As an alternative to Kernel PCA, PCA was used, but did not lead to better results. Kernel PCA increased the number of input features to 300. The following values obtained are worse than those obtained in the test in which the input samples are selected according to standard deviation. The average of the Score for each soil parameter is 1.133.

The same test is repeated using 1200 samples for training data, from the initial dataset, where values are not selected according to standard deviation. Similar values are obtained. All the data results are linked in Table 7. The reduction in the number of samples prevents DL models from functioning optimally, as neural networks usually need thousands of samples to function properly. In this case, sample selection based on the standard deviation value, reduced the number of samples to less than 100. Testing with the previously used machine learning algorithms yielded the best results with the RF algorithm.

A further test (ablation study) a RF model, using only three of the four output parameters (P_2O_5, K, Mg), with a skimmed training dataset containing 60 samples with the lowest standard deviation values yields the following values of MSE, with a Final Score: 0.905 (Table 8).

The feature graphs in Fig. 6-8 show in green the ground truth, in red the baseline, in blue the prediction.

	P_2O_5	K	Mg	pH
MSE	564.095	5642.910	766.010	708.752 e-02
MSE baseline	552.402	3669.063	788.853	705.276 e-02
Score	1.021	1.537	0.971	1.004

Table 6. Results of 70 samples without taking to account the standard deviation

	P_2O_5	K	Mg	pH
MSE	859.110	3645.352	1353.572	1.11391e-01
MSE baseline	781,619	3391,023	1278.887	1.11725e-01
Score	1.099	1.075	1.051	0.997

Table 7. Results of 1200 samples without taking to account the standard deviation

	P_2O_5	K	Mg
MSE	426.937	2490.950	1147.269
MSE baseline	452.515	2647.150	1377.112
Score	0.943	0.940	0.833

Table 8. Random Forest model score

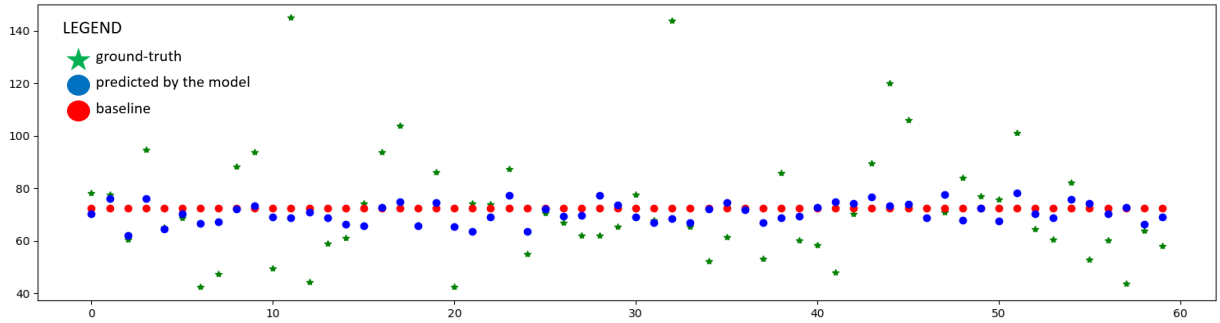

Figure 6. The values predicted by Authors’ approach for the output parameters P_2O_5.

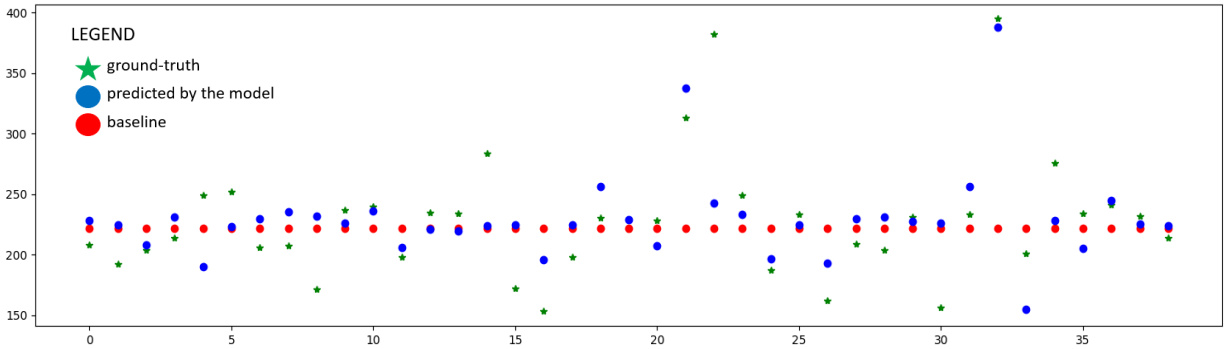

Figure 7. The values predicted by Authors’ approach for the output parameters K.

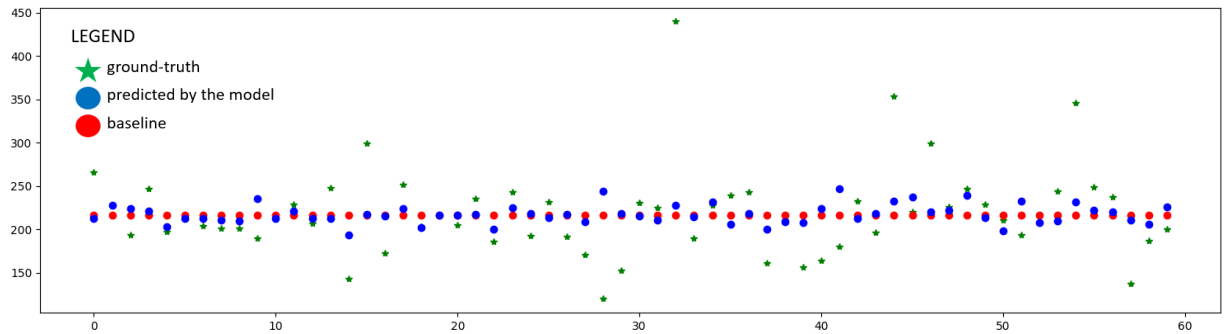

Figure 8. The values predicted Authors’ approach for the output parameters Mg.

4. Conclusions

4.1.1 To cope with the huge problems of climate change and insufficient production compared to the exponential urban growth we are witnessing, the possibility of coupling agriculture with technology becomes necessary. It is therefore a primary need to have continuous knowledge of the concentration of nutrients in the soil. It passes from areas where there is no production due to the absence of these nutrients, to areas where there is excessive crop production depending on an increase in soil pollution due to fertilizers. A solution should therefore be found to help agriculture in better crop management and constant soil monitoring. The best solution from an economic and logistical point of view is offered using remote sensing and machine learning techniques. This is important also because using the traditional monitoring of the soil by the extraction of samples in the crop means that, then, in the laboratory it is necessary to use other chemicals to exploit the presence of the parameter to be detected. It is nowadays important to advance a sustainable monitoring idea by reducing the on-site inspections and guarantee a sustainable management of natural resources (Michałowska et al., 2022). In accordance with the reasons expressed above, a hyperspectral sensor would be able to estimate the presence of parameters in the soil by avoiding on-site inspections. Acquisition made by a hyperspectral sensor applied on a UAV device requires less time and covers a larger survey area. The estimation of soil parameters by hyperspectral images is a big challenge that would bring a breakthrough in sustainable agriculture. In that paper the dataset was composed by extensive ground samplings collocated with airborne hyperspectral measurements from imagers mounted on Piper PA-31 Navajo aircraft. The aim of the challenge was to automatically estimate the soil parameters (K, P_2O_5, Mg and pH) using artificial intelligence techniques. The difference in the method presented here respect to the others lets an improvement in metrics in the model performance. The novelty of the authors’ is in the approach for this educational challenge because it is based on the skimming of the input dataset selecting all the data with the lower dispersion. The scores achieved with this methodology result to be poorest respect to the scores achieved with the other performed models with an unskimmed input dataset. Other factors that influenced negatively the accuracy of the model have been the size of the patch and the number of the bands. According to all that troubles, the authors of this paper considered the skimming phase as the best hypothesis to improve the model performance. Using a larger number of samples characterized by low data dispersion, the use of a neural network such as the one described in that work, could be regarded as an alternative to the classical models used. This requires a dataset containing several tens of thousands of carefully selected samples. As future approach can be useful having a higher number of skimmed samples to test other neural networks.

Acknowledgements

The authors thank the AI4EO Challenge, organized by the European Space Agency (ESA), for making available the dataset used in this study.

Fundings

This research is supported by the European Union - Next Generation EU through the Project of National Interest (PRIN) "Geo-Intelligence for improved air quality monitoring and analysis (GeoAIr)" 202258ACSL - PE10 at the Italian *Ministry for University and Research* (*MUR*).

References

Balasundram, S.K., Shamshiri, R.R., Sridhara, S., Rizan, N., 2023: The Role of Digital Agriculture in Mitigating Climate Change and Ensuring Food Security: An Overview, *Sustainability*, 15(6), 5325. doi.org/10.3390/su15065325.

Bibi, F., Rahman, A., 2023: An Overview of Climate Change Impacts on Agriculture and Their Mitigation Strategies, *Agriculture*, 13(8), 1508. doi.org/10.3390/agriculture13081508.

Chiappini, S., et al., 2023: Time series analysis of olive orchard coverage in the rural landscape: a case study of the Cartoceto Municipality, 2023 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Pisa, Italy, pp. 746-751, doi.org/10.1109/MetroAgriFor58484.2023.10424269.

Dhiman, G., Bhattacharya, J., Roy, S., 2023: Soil textures and nutrients estimation using remote sensing data in north india - Punjab region, *Procedia Computer Science*, 218, 2041-2048. doi.org/10.1016/j.procs.2023.01.180.

Filho, W.L., Nagy, G.J., Setti, A.F.F., Sharifi, A., Donkor, F.K., Batista, K., Djekic, I., 2023: Handling the impacts of climate change on soil biodiversity, *Science of The Total Environment*, 869, 2023, 161671. doi.org/10.1016/j.scitotenv.2023.161671.

ASVIS, Goal e Target: obiettivi e traguardi per il 2030, [online:] https://asvis.it/goal-e-target-obiettivi-e-traguardi-per-il-2030/ [access: 15.07.2024].

Ghosh, A., Kumar, A., Biswas, G., 2024: Chapter 1 - Exponential population growth and global food security: challenges and alternatives, Editor(s): Prasann Kumar, Arun Lal Srivastav, Veena Chaudhary, Eric D. van Hullebusch, Rosa Busquets, Bioremediation of Emerging Contaminants from Soils, Elsevier, 1-20, ISBN 9780443139932, doi.org/10.1016/B978-0-443-13993-2.00001-3.

Hossain, M.D., Kashem, M.A., Mustary, S., 2023: IoT Based Smart Soil Fertilizer Monitoring And ML Based Crop Recommendation System, 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), Chittagong, Bangladesh, pp. 1-6, doi.org/10.1109/ECCE57851.2023.10100744.

Hosseini, F.S., Seo, M.B., Razavi-Termeh, S.V., Sadeghi-Niaraki, A., Jamshidi, M., Choi, S.-M., 2023: Geospatial Artificial Intelligence (GeoAI) and Satellite Imagery Fusion for Soil Physical Property Predicting, *Sustainability*, 15(19), 14125. doi.org/10.3390/su151914125.

Khanal, S., KC, K., Fulton, J.P., Shearer, S., Ozkan, E., 2020: Remote Sensing in Agriculture—Accomplishments, Limitations, and Opportunities, *Remote Sensing*, 12(22), 3783. doi.org/10.3390/rs12223783.

KP LABS, Intuition-1. Available, [online:] https://kplabs.space/intuition-1/ [access: 10/07/2024]

Kuzu, R.S., Albrecht, F., Arnold, C., Kamath, R., Konen, K., 2022: Predicting Soil Properties from Hyperspectral Satellite Images, 2022 IEEE International Conference on Image

Processing (ICIP), Bordeaux, France, pp. 4296-4300, doi.org/10.1109/ICIP46576.2022.9897254.

Michałowska, K., Pirowski, T., Głowienka, E., Szypuła, B., Malinverni, E.S., 2024: Sustainable Monitoring of Mining Activities: Decision-Making Model Using Spectral Indexes, *Remote Sensing*, 16(2), 388. doi.org/10.3390/rs16020388.

Moersdorf, J., Rivers, M., Denkenberger, D., Breuer, L., Jehn, F.U., 2024: The Fragile State of Industrial Agriculture: Estimating Crop Yield Reductions in a Global Catastrophic Infrastructure Loss Scenario, *Global Challenges*, 8(1), 2300206. doi.org/10.1002/gch2.202300206.

Mukhamediev, R.I., Merembayev, T., Kuchin, Y., Malakhov, D., Zaitseva, E., Levashenko, V., Popova, Y., Symagulov, A., Sagatdinova, G., Amirgaliyev, Y., 2023: Soil Salinity Estimation for South Kazakhstan Based on SAR Sentinel-1 and Landsat-8,9 OLI Data with Machine Learning Models, *Remote Sensing*, 15(17), 4269. doi.org/10.3390/rs15174269.

Nalepa, J., et al., 2024: Estimating Soil Parameters From Hyperspectral Images: A benchmark dataset and the outcome of the HYPERVIEW challenge, *IEEE Geoscience and Remote Sensing Magazine*, 12(3), 35-63. doi.org/10.1109/MGRS.2024.3394040.

Nalepa, J., et al., 2022: The Hyperview Challenge: Estimating Soil Parameters from Hyperspectral Images, 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, pp. 4268-4272, doi.org/10.1109/ICIP46576.2022.9897443.

Ocwa, A., Harsanyi, E., Széles, A., Holb, I.J., Szabó, S., Rátonyi, T., Mohammed, S., 2023: A bibliographic review of climate change and fertilization as the main drivers of maize yield: implications for food security, *Agriculture & Food Security*, 12. doi.org/10.1186/s40066-023-00419-3.

Pande, C.B., Moharir, K.N., 2023: Application of Hyperspectral Remote Sensing Role in Precision Farming and Sustainable Agriculture Under Climate Change: A Review. In: Pande, C.B., Moharir, K.N., Singh, S.K., Pham, Q.B., Elbeltagi, A. (eds) Climate Change Impacts on Natural Resources, Ecosystems and Agricultural Systems. Springer Climate. Springer, Cham. doi.org/10.1007/978-3-031-19059-9_21.

Pandey, P. C., Pandey, M., 2023: Highlighting the role of agriculture and geospatial technology in food security and sustainable development goals, *Sustainable Development*, 31(5), 3175-3195. doi.org/10.1002/sd.2600.

Penuelas, J., Coello, F., Sardans, J., 2023: A better use of fertilizers is needed for global food security and environmental sustainability, *Agriculture & Food Security*, 12. doi.org/10.1186/s40066-023-00409-5.

The European Space Agency, Platform AI4EO challenge Seeing beyond the visible https://platform.ai4eo.eu/seeing-beyond-the-visible/data (accessed on 10/07/2024)

Santurkar. S., Tsipras. D., Ilyas, A., Madry A., 2018: How Does Batch Normalization Help Optimization? In Proceedings of 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada, doi.org/10.48550/arXiv.1805.11604.

Sarfraz, S., Ali, F., Hameed, A., Ahmad, Z., Riaz, K., 2023: Sustainable Agriculture Through Technological Innovations. In: Prakash, C.S., Fiaz, S., Nadeem, M.A., Baloch, F.S., Qayyum, A. (eds) Sustainable Agriculture in the Era of the OMICs Revolution. Springer, Cham. doi.org/10.1007/978-3-031-15568-0_10.

Santosh, D.T., et al., 2023: Alleviation of Climate Catastrophe in Agriculture Through Adoption of Climate-Smart Technologies. In: Chatterjee, U., Shaw, R., Kumar, S., Raj, A.D., Das, S. (eds) Climate Crisis: Adaptive Approaches and Sustainability. Sustainable Development Goals Series. Springer, Cham. doi.org/10.1007/978-3-031-44397-8_17.

Song, Y., Ye, M., Zheng, Z., Zhan, D., Duan, W., Lu, M., Song, Z., Sun, D., Yao, K., Ding, Z., 2023: Tree-Structured Parzan Estimator–Machine Learning–Ordinary Kriging: An Integration Method for Soil Ammonia Spatial Prediction in the Typical Cropland of Chinese Yellow River Delta with Sentinel-2 Remote Sensing Image and Air Quality Data, *Remote Sensing*, 15(17), 4268. doi.org/10.3390/rs15174268.

Zayani, H., Fouad, Y., Michot, D., Kassouk, Z., Baghdadi, N., Vaudour, E., Lili-Chabaane, Z., Walter, C., 2023: Using Machine-Learning Algorithms to Predict Soil Organic Carbon Content from Combined Remote Sensing Imagery and Laboratory Vis-NIR Spectral Datasets, *Remote Sensing*, 15(17), 4264. doi.org/10.3390/rs15174264.

Zelenka, C., Lohrer, A., Bayer, M., Kröger, P., 2022: AI4EO Hyperview: A Spectralnet3d and Rnnplus Approach for Sustainable Soil Parameter Estimation on Hyperspectral Image Data, 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, pp. 4263-4267, doi.org/10.1109/ICIP46576.2022.9897889.

Yao, L.; Xu, M.; Liu, Y.; Niu, R.; Wu, X.; Song, Y., 2024: Estimating of heavy metal concentration in agricultural soils from hyperspectral satellite sensor imagery: Considering the sources and migration pathways of pollutants, *Ecological Indicators*, 158, 111416. doi.org/10.1016/j.ecolind.2023.111416.

Yu, B., Yuan, J., Yan, C., Xu, J., Ma, C., Dai, H., 2023: Impact of Spectral Resolution and Signal-to-Noise Ratio in Vis–NIR Spectrometry on Soil Organic Matter Estimation, *Remote Sensing*, 15(18), 4623. doi.org/10.3390/rs15184623.