IOCAG 2022

MDPI

Proceeding Paper

Crop Identification by Machine Learning Algorithm and Sentinel-2 Data †

Serafeim Stournaras [1,*], Dimitrios Loukatos [2], Konstantinos Arvanitis [2] and Nikos Kalatzis [3]

[1] Department of Natural Resources Management and Agricultural Engineering, Agricultural University of Athens, 75 Iera Odos Str., 118 55 Athens, Greece; stud115002@aua.gr and srf.stour@gmail.com
[2] , Department of Natural Resources Management and Agricultural Engineering, Agricultural University of Athens, 75 Iera Odos Str., 118 55 Athens, Greece; {dlouka; karvan}@aua.gr
[3] Computer software engineering, Technical Project Manager in Neuropublic S.A., Methonis 6 Str., 185 45 Piraeus, Greece; n_kalatzis@neuropublic.gr
* Correspondence: stud115002@aua.gr; srf.stour@gmail.com
† Presented at the 1st International Online Conference on Agriculture—Advances in Agricultural Science and Technology (IOCAG2022), 10–25 February 2022; Available online: https://iocag2022.sciforum.net/.

Abstract: There is a growing need for remote identification of the crop types, which is a serious issue for policy makers and statistical accountants (i.e., agricultural inspectors and government agencies), for verifying the degree of validity of the information concerning the area and the type of each crop being cultivated. In this work, remote Sentinel-2 imaging data was utilized for calculating average NDVI values, twice a month, for the period 2017–2020, for cotton, rice and olive trees. In addition, a machine learning algorithm was developed and the corresponding model was trained using the average NDVI values. Python programming and KNN machine learning on the PyCharm environment were used.

Keywords: crop identification; NDVI; Sentinel-2; machine learning

Citation: Stournaras, S.; Loukatos, D.; Arvanitis, K.; Kalatzis, N. Crop Identification by Machine Learning Algorithm and Sentinel-2 Data. *Chem. Proc.* **2022**, *3*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s):

Published: date

1. Introduction

The degradation of the arable land and the water resources and the rapid growth of world population intensify the need for enhancing the agricultural productivity [1]. The latter goal is facilitated by the accurate mapping and crop-type identification for supporting crop growth monitor, yield prediction, and global food-security decisions [2]. Spatial information on the distribution of arable land, in conjunction with crop type identification, can help in valuable and accurate statistical estimations, such as forecasting of agricultural production and crop area estimation, thus improving the efficiency of agricultural policy mechanisms. Accurate crop maps, generated by sensors, by remote earth observation, or by a combination of them, can form the basis for agricultural monitoring and decision-making in remote areas, to support a sustainable agricultural land management [3]. Due to the fact that each of crop has its unique phenology and that the phenological differences among the crops can be described by time series of remotely sensed images [4], the satellite time-series images have been widely used for annual crop classification [3,5]. Most existing annual crop classifications use the image time series of the entire growing season to identify crop [5,6].

Indeed, remote sensing data has found its way into precision agriculture with the aim of increasing agricultural efficiency. Second, remote sensing is a valuable tool for monitoring agricultural expansion. Finally, it provides timely, comprehensive, objective, transparent, accurate and non-discriminatory data, where the resulting remote information can be used without hesitation. Most existing annual crop classifications use the image time series of the entire growing season to identify crop [3].

Furthermore, as NDVI time series can be used to describe the phenological differences among different crops, Hao et al. [7] propose a NDVI time series-based method

(RBM). This method generates reference NDVI time series that can be used to identify crop types and produce training samples [8]. It is generally accepted that phenological and growing patterns of a crop are actual similar or the same in the different regions of the world [9]. Based on this phenomenon, it is reasonable to hypothesize that a supervised classification model trained in one region can be applied to other regions to identify the crops common in both the training and applied regions. Since major types of crops, like olive trees, wheat and rice are distributed globally, it will make the global in-season identifying of major crops possible. The key element is to train a classification model efficient enough to compensate the differences between the actual crop growth environment and the training region where plentiful ground samples are openly available, without sacrificing prediction accuracy [10,11].

Benefited from the rapid growth of the ICT (Information and Communication Technologies) technologies that provide well-documented and easy-to-use programming environments [12], this work presents a simple crop identification method, for three different crop types, using machine learning, which exploits growing season's NDVI time series extracted by Sentinel-2 (S2) data. These data were acquired over the area of interest, from the beginning of growing season in 2017 until the end of year 2020, were further processed, compared against ground-truth data of cover crops, and fed machine learning techniques in python environment.

2. Methods and materials

Ground-based data were collected by Sentinel-2 (S2) at most of times fortnightly during of the year. The three crop types selected were cotton, rice and olive trees. Data availability and validity were assured by Neuropublic SA [http://www.neuropublic.gr] company. Sentinel-2 is a wide-swath, high-resolution, multi-spectral imaging mission, supporting Copernicus Land Monitoring studies, including the monitoring of vegetation, soil and water cover, as well as observation of inland waterways and coastal areas. The Sentinel-2 Multispectral Instrument (MSI) collects data using 13 spectral bands, with four bands at 10 m, six bands at 20 m and three bands at 60 m of spatial resolution [13].

In 2021, S2 data covering the spatial distribution of the training samples selected in this study were utilized for each crop type. The digital data utilized to calculate the NDVI values were based on Band 4 (0.665 μm) and Band 8 (0.842 μm) for the period 2017–2020. We collected S2 data between day of year 01 and 365, then calculated NDVI for each image with Red and NIR bands [14], and then generated NDVI time series by selecting the maximum NDVI value within each 15-day window. The 15-day time granularity was necessary because cloud-free images cannot be acquired daily, while an every 15-day image time series could best describe the crop phenological difference and reduce the number of missing values, which is the best choice for in-season crop classification [15].

In case that there was no value to be assigned as a 15-day sample, a gap was marked and the 15-day composited NDVI time series were initially filled via a moving window method, by calculating the average of the two neighboring high-quality values in the time series, and finally smoothed using Savitzky-Golay (S-G) filters, to further ameliorate irregular variations in the NDVI time series [3,16]. The equation for calculating NDVI is given by Equation (1).

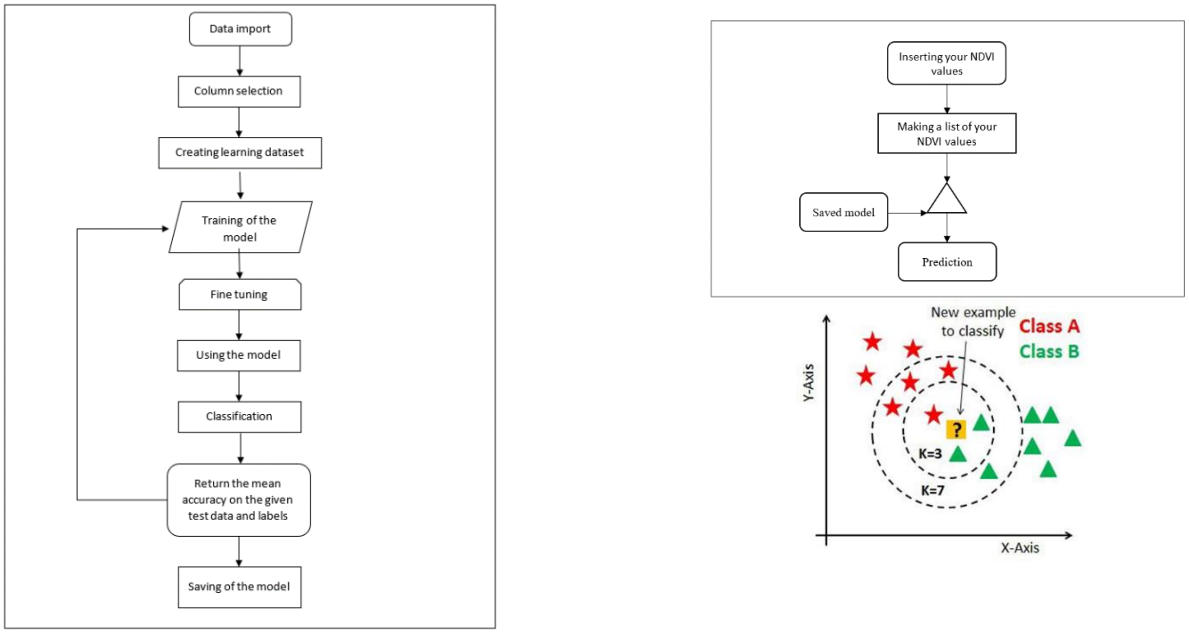
$$\text{NDVI}=\frac{\rho(NIR)-\rho(Red)}{\rho(NIR)+\rho(Red)} \quad (1)$$

Equation (1): for calculating NDVI.

Where ϱ(NIR) and ϱ(Red) donate the SR reflectance of NIR and Red band respectively, which are Band 4 and Band 8 of Sentinel-2 data. Afterwards, the aggregate NDVI time series of all these training samples were composed from the separate 15-day NDVI image time series between 2017 and 2020. All values were collected in one MS Excel datasheet file, for further processing.

Both the trained algorithm (Figure 1a) and identification algorithm (Figure 1b: top) were implemented in python language. Figure 1a depicts the algorithm that has been used

for training and testing the model. The preprocessed data were imported into algorithm by MS excel datasheets that were generated from the averaged values of NDVI time series. Some of the selected columns and series create a learning dataset for the training of its own and the others used for checking of the training. We have to choose the columns that presented raises of the NDVI values because correspond at increase of plant's biomass and helping future predictions. Trained algorithm uses 'pandas' module for the data entry and scikit-learn module with KNN classifier about machine learning. At the end of Figure 1a, the command 'saving of the model', in particular, generates a .joblib file that is going to be used by the second algorithm (Figure 1b: top), thus providing crop type identification based on user-provided data [17].



(**a**) (**b**)

Figure 1. (**a**) The analysis of the training algorithm; (**b**) top: Identification of crop type by machine learning algorithm, bottom: KNN module on figure (source: https://medium.com/).

The top part of Figure 1b explains the identification algorithm that makes the prediction of the crop type. Identification algorithm asks as input the NDVI values referring to a specific period and uses scikit-learn [17,18] module for rerunning the saved model. These input NDVI values depend on cultivation season and are connected with the plant's biomass. After that, the algorithm fits the values being inserted in the already trained algorithm (i.e., the model) and thus makes the identification of the crop type.

For the training of the algorithm is used the K-Nearest Neighbors (KNN) module (Figure 1.b: bottom). K-Nearest Neighbors (KNN) is a type of supervised learning algorithm used for both regression and classification. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then closest to the test data points selected as a K number. The KNN algorithm calculates the probability of the test data belonging to the classes of 'K' training data and class holds the highest probability will be selected. In the case of regression, the value is the mean of the 'K' selected training points [18,19].

KNN Algorithm Description:

1. Select the K number of the neighbors. K value indicates the count of the nearest neighbors;
2. Calculate the Euclidean distance of K number of neighbors;
3. Take the K nearest neighbors as per the calculated Euclidean distance;
4. Among these k neighbors, count the number of the data points in each category;

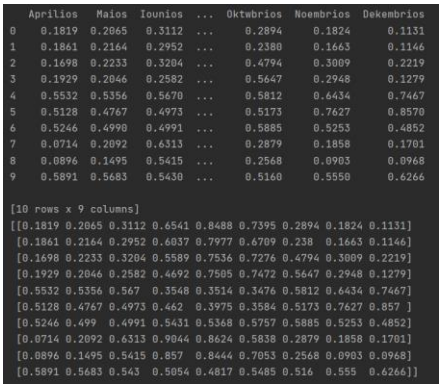
5. Assign the new data points to that category for which the number of the neighbor is maximum;
6. The KNN model is ready.

QGIS is free and open-source platform desktop geographic information system (GIS) application that supports graphic viewing and editing of geospatial data. In this paper, data were utilized with QGIS Desktop 3.16.2 with GRASS 7.8.4 that allowing user to analyze and edit geospatial data, in order to export a graphical map. Graphical maps were exported and showed the heterogynous of the crop types due to different growing and climate conditions, but also shows the heterogynous between different and same crop's phenological stages by collected NDVI.

3. Experimentation, Results and Discussion

3.1. Training Process

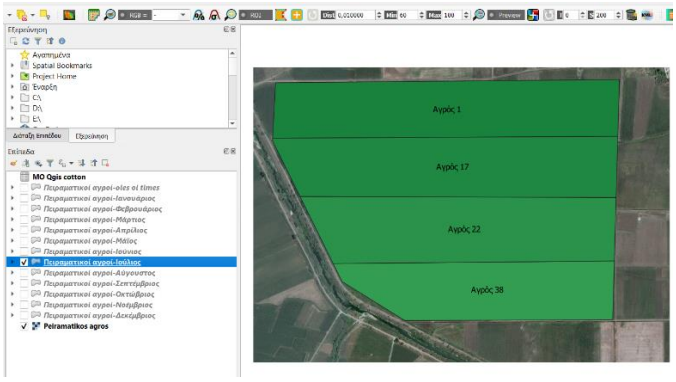
Three different crops were analyzed: cotton, rice and olive tree. The data used for the experiment were preprocessed and monthly average of NDVI values were calculated for each month. Preprocess included the typical method of average. Next, a machine learning algorithm was developed and training was accomplished utilizing monthly average NDVI values. Python programming language and KNN machine learning module (Figure 2a) on a PyCharm shell were used for the development of the machine learning algorithm. NDVI data that concentrated in period 2017 until 2020 (Figure 2c) were utilized by the algorithm on eighty (80) percent for training and twenty (20) percent for testing its own. Figure 2b,d provide the graphical presentation of cotton and olive tree, respectively, NDVI datasets, for month July, using color indicators. For verification purposes used QGIS program.



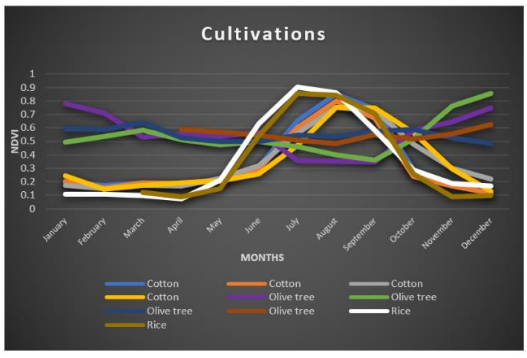
```
   Aprilios   Maios  Iounios  ...  Oktwbrios  Noembrios  Dekembrios
0    0.1819  0.2065   0.3112  ...     0.2894     0.1824      0.1131
1    0.1861  0.2164   0.2952  ...     0.2380     0.1663      0.1146
2    0.1698  0.2233   0.3204  ...     0.4794     0.3009      0.2219
3    0.1929  0.2046   0.2582  ...     0.5647     0.2948      0.1279
4    0.5532  0.5356   0.5670  ...     0.5812     0.6434      0.7467
5    0.5128  0.4767   0.4973  ...     0.5173     0.7627      0.8570
6    0.5246  0.4990   0.4991  ...     0.5885     0.5253      0.4852
7    0.0714  0.2092   0.6313  ...     0.2879     0.1858      0.1701
8    0.0896  0.1495   0.5415  ...     0.2568     0.0903      0.0968
9    0.5891  0.5683   0.5430  ...     0.5160     0.5550      0.6266

[10 rows x 9 columns]
[[0.1819 0.2065 0.3112 0.6541 0.8488 0.7395 0.2894 0.1824 0.1131]
 [0.1861 0.2164 0.2952 0.6037 0.7977 0.6709 0.238  0.1663 0.1146]
 [0.1698 0.2233 0.3204 0.5589 0.7536 0.7276 0.4794 0.3009 0.2219]
 [0.1929 0.2046 0.2582 0.4692 0.7505 0.7472 0.5647 0.2948 0.1279]
 [0.5532 0.5356 0.567  0.3548 0.3514 0.3476 0.5812 0.6434 0.7467]
 [0.5128 0.4767 0.4973 0.462  0.3975 0.3584 0.5173 0.7627 0.857 ]
 [0.5246 0.499  0.4991 0.5431 0.5368 0.5757 0.5885 0.5253 0.4852]
 [0.0714 0.2092 0.6313 0.9044 0.8624 0.5838 0.2879 0.1858 0.1701]
 [0.0896 0.1495 0.5415 0.857  0.8444 0.7053 0.2568 0.0903 0.0968]
 [0.5891 0.5683 0.543  0.5054 0.4817 0.5485 0.516  0.555  0.6266]]
```

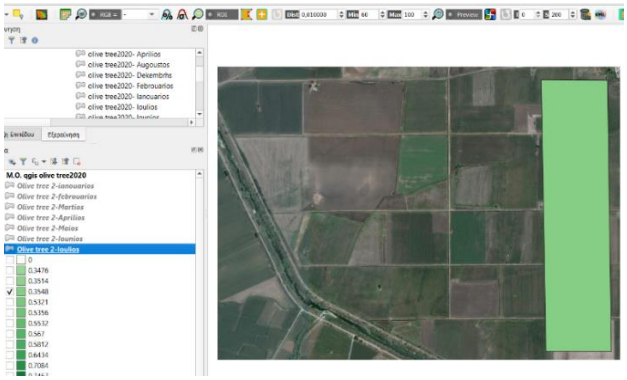
(**a**)



(**b**)



(**c**)



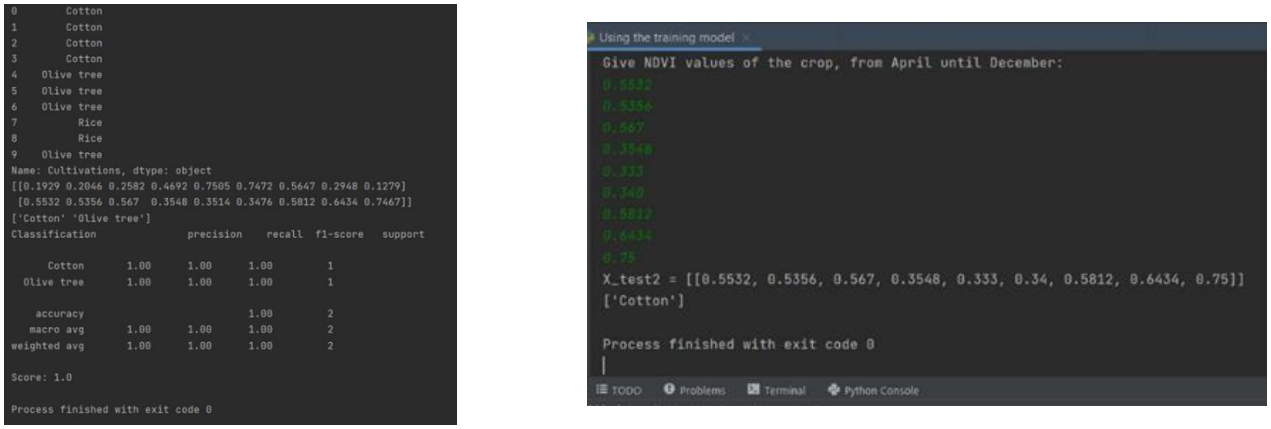
(**d**)

Figure 2. (**a**) Run of the training algorithm; (**b**) Graphical presentation cotton crops' NDVI dataset on month July using color indicator; (**c**) Figure of training NDVI dataset; (**d**) Graphical presentation olive tree crop's NDVI dataset on month July using color indicator.

The trained model was saved on a .joblib file. As it was anticipated, after two or three runs the algorithm's accuracy has become '1.00', because of the low number of the samples that are on the training NDVI dataset, Figure 3a. The fast convergence of the training model process is attribute to the limited amount of the original data being in used.

3.2. Identification Process

The algorithm that was developed for identifying the crops was tested by inserting used defined NDVI series values and comparing them against the trained algorithm (.joblib). Through this process, the entry level identification of crop types was accomplished based on few NDVI values. The trained model algorithm implementation (Figure 3b) asks user for specific NDVI values, per month, for a specific period and cultivation. Finally, the result, i.e., the matching crop type being identified, is provided by a textual output as depicted in the bottom part of Figure 3b.



(**a**) (**b**)

Figure 3. (**a**) Testing results by trained machine learning algorithm; (**b**) Input for the unknown crop's NDVI values and identification prediction.

In this case, user-given values are colored in green (Figure 3b). The execution of the testing algorithm using slightly different input values returned the same (crop type) result. These results where fast-generated and satisfactory, for this simple machine learning implementation.

3.3. Discussion

In this paper, a method has been presented for introducing the potential of machine learning techniques using python for crop identification purposes. This approach is beneficial for either students of agriculture or professionals wanting to become familiar with the techniques of the digital era. The dataset being used was quite limited, but further ongoing research, combining this method with richer NDVI timeseries, is delivering satisfactory results that will be included in a more mature version of this preliminary work. Apart from the core python-based machine learning engine, the role of assistive open-source tools for elaboration and visualization of geospatial agricultural-specific data, like the QGIS, is also highlighted.

4. Conclusions

This paper highlighted the fusibility of implementing, presented a simple K-nearest neighbor (KNN) model, in which classification models were trained with a type of supervised learning algorithm used for both classification and regression. KNN tries to predict the correct crop type from the test data by calculating the distance between the test data and all the training points corresponding to composited NDVI time series. Training NDVI data were collected across the Greece to contain NDVI time series of each crop under different climate and irrigation conditions. The trained classification model was then tested

for crop identification. The performance of this KNN model was tested using real data. The learning method achieved proper identification results when using NDVI time series referring to the entire growing season. The identification of the crop type by slightly different NDVI values was satisfactory. Training, refinements and tests using more data, as well as better visualization of the results, will be significant future objectives.

Acknowledgments: The authors are grateful to the company Neuropublic S.A. for providing access to the original crop data that were used in this study.

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement:

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

References

1. Bajželj, B.; Richards, K.S.; Allwood, J.M.; Smith, P.; Dennis, J.S.; Curmi, E.; Gilligan, C.A. Importance of food-demand management for climate mitigation. *Nat. Clim. Change* **2014**, *4*, 924–929.
2. Lobell, D.B. The use of satellite data for crop yield gap analysis. *Field Crops Res.* **2013**, *143*, 56–64.
3. Hao, P.; Di, L.; Zhang, C.; Guo, L. Transfer Learning for Crop classification with Cropland Data Layer data (CDL) as training samples. *Sci. Total Environ.* **2020**, *733*, 138869.
4. Zhang, J.; Feng, L.; Yao, F. Improved maize cultivated area estimation over a large scale combining MODIS-EVI time series data and crop phenological information. *ISPRS J. Photogramm. Remote Sens.* **2014**, *94*, 102–113.
5. Zhong, L.; Hu, L.; Zhou, H. Deep learning based multi-temporal crop classification. *Remote Sens. Environ.* **2019**, *221*, 430–443.
6. Löw, F.; Michel, U.; Dech, S.; Conrad, C. Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using support vector machines. *ISPRS J. Photogramm. Remote Sens.* **2013**, *85*, 102–119.
7. Hao, P.; Wang, L.; Zhan, Y.; Niu, Z. Using moderate-resolution temporal NDVI profiles for high-resolution crop mapping in years of absent ground reference data: a case study of bole and Manas Counties in Xinjiang, China. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 67.
8. Hao, P.; Wang, L.; Zhan, Y.; Wang, C.; Niu, Z.; Wu, M. Crop classification using crop knowledge of the previous year: case study in Southwest Kansas, USA. *Eur. J. Remote Sens.* **2016**, *49*,1061–1077.
9. Zhong, L.; Gong, P.; Biging, G.S. Efficient corn and soybean mapping with temporal extendability: A multi-year experiment using Landsat imagery. *Remote Sens. Environ.* **2014**, *140*, 1–13.
10. Boryan, C.; Yang, Z.W.; Mueller, R.; Craig, M. Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto Int.* **2011**, *26*, 341–358.
11. Han, W.; Yang, Z.; Di, L.; Zhang, B.; Peng, C. Enhancing agricultural geospatial data dissemination and applications using geospatial web services. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4539–4547.
12. Arvanitis, K.G.; Symeonaki, E.G. Agriculture 4.0: The Role of Innovative Smart Technologies Towards Sustainable Farm Management. *Open Agric. J.* **2020**, *14*, 130–136.
13. Sentinel Online. Available online: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/overview (accessed on 20 December 2021).
14. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote *Sens. Environ.* **2002**, *83*, 195–213.
15. Hao, P.; Wu, M.; Niu, Z.; Wang, L.; Zhan, Y. Estimation of different data compositions for early-season crop type classification. *PeerJ* **2018**, *6*, e4834.
16. Zhang, X.; Liu, L.; Liu, Y.; Jayavelu, S.; Wang, J.; Moon, M.; Henebry, G.M.; Friedl, M.A.; Schaaf, C.B. Generation and evaluation of the VIIRS land surface phenology product. *Remote Sens. Environ.* **2018**, *216*, 212–229.
17. Model Persistence. Available online: https://scikit-learn.org/stable/modules/model_persistence.html (accessed on 20 December 2021).
18. Nearest Neighbors. Available online: https://scikit-learn.org/stable/modules/neighbors.html (accessed on 20 December 2021).
19. Patwardhan Sai. Simple Understanding and Implementation of KNN Algorithm. https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/ (accessed on 20 December 2021).