Heliyon 9 (2023) e13339

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Investigation on the use of ensemble learning and big data in crop identification

Sayed Ahmed [a], Amira S. Mahmoud [a], Eslam Farg [a], Amany M. Mohamed [a], Marwa S. Moustafa [a], Khaled Abutaleb [a], Ahmed M. Saleh [a], Mohamed A.E. AbdelRahman [a,*], Hisham M. AbdelSalam [b], Sayed M. Arafat [a]

[a] *National Authority for Remote Sensing and Space Science (NARSS), Cairo, Egypt*
[b] *Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt*

ARTICLE INFO

Keywords:
Big data
Crop identification
Ensemble learning
DB Framework
Apache spark

ABSTRACT

The agriculture sector in Egypt faces several problems, such as climate change, water storage, and yield variability. The comprehensive capabilities of Big Data (BD) can help in tackling the uncertainty of food supply occurs due to several factors such as soil erosion, water pollution, climate change, socio-cultural growth, governmental regulations, and market fluctuations. Crop identification and monitoring plays a vital role in modern agriculture. Although several machine learning models have been utilized in identifying crops, the performance of ensemble learning has not been investigated extensively. The massive volume of satellite imageries has been established as a big data problem forcing to deploy the proposed solution using big data technologies to manage, store, analyze, and visualize satellite data. In this paper, we have developed a weighted voting mechanism for improving crop classification performance in a large scale, based on ensemble learning and big data schema. Built upon Apache Spark, the popular DB Framework, the proposed approach was tested on El Salheya, Ismaili governate. The proposed ensemble approach boosted accuracy by 6.5%, 1.9%, 4.4%, 4.9%, 4.7% in precision, recall, F-score, Overall Accuracy (OA), and Matthews correlation coefficient (MCC) metrics respectively. Our findings confirm the generalization of the proposed crop identification approach at a large-scale setting.

1. Introduction

Smart farming is emphasized in Egypt's data-driven economic reform [1] affected by soil characteristics, water availability, and harvesting practices [2]. Crop discrimination is essential in developing smart farming systems that helps facilitate crop management and yield forecasting [3]. Remote sensing sensors [4] (optical and microwave) were favored in terms of cost and time in crop management and yield forecasting [5]. Understanding the characteristic of electromagnetic wavelength behavior of crop is essential in crop identification [6]. The electromagnetic response of a crop cover depends not only on wavelength [7], but also on season, sensor angle, crop status, illumination intensity, weather phenomenon and topography among other external factors. According to Ref. [8] Crop coefficient is varying according to growth stage and also affected by the growth stage length. The analysis shows also that Normalized Difference Vegetation Index (NDVI), SoilAdjusted Vegetation Index (SAVI), crop coefficient (Kc) and predicted Kc had the same trend

* Corresponding author.
E-mail addresses: maekaoud@gmail.com, maekaoud@yahoo.com, maekaoud@narss.sci.eg (M.A.E. AbdelRahman).

https://doi.org/10.1016/j.heliyon.2023.e13339
Received 14 June 2022; Received in revised form 4 January 2023; Accepted 25 January 2023
Available online 31 January 2023

through the different growth stages. Also different approaches combines different remote sensing data sources applied for cropland mapping [9]. developed an approach based on GEE for accurate tree-fruits mapping by testing different temporal stacking windows, spectral stacking methods, and various integration scenarios between Sentinel-2 optical (S2) and Sentinel-1 SAR (S1) data as inputs to the Random Forest (RF) classifier. Comparative accuracy analysis showed that using time series S2 spectral bands (SBs) and S1 polarization channels with some added S1 textural features could use as best integration scenario that achieved the highest accuracy (OA = 96.31 & Kappa = 0.96) was adding S1 textural features to S2 spectral features. In addition to [10], used time series of satellite images calculated Vegetation Index (VI)'s the results showed that the best representation of the crop phenological changes during the crop growth season and higher accuracy in strategic crops discrimination with overall kappa accuracy with 0.82 and 0.79 respectively. On the other hand, shortage of processing power and huge amount of data were a challenge for proposed method to apply on all Egypt.

Microwave sensors can penetrate clouds and vegetation better than optical sensors achieving a boost in crop discrimination [11]. Recent studies included physical and handcrafted features. Physical features include NDVI [12] and Leaf Area Index (LAI) [13] used time series of satellite images calculated VI's the results showed that the best representation of the crop phenological changes during the crop growth season and higher accuracy in strategic crops discrimination.

Handcrafted features include orthogonal subspace projections, Principal Component Analysis (PCA) [14], and Minimum Noise Fraction (MNF) [14]. The huge volumes of satellite imageries had to be processed increase the need to big data technologies to be incorporated in agriculture problems [15]. Using remote sensing helps in an accurate crop inventory under complex landscape conditions based on the spectral characteristics of differences crops. That because the agricultural fields in Egypt are commonly distributed with relatively small sizes parcels, which usually reduce the reliability of Agricultural statistics in surveying cropland [8].

The available land use/cover datasets have only one cropland category, with no detailed information on crop types, areas, and spatial distribution, which are essential information for a wide range of agriculture applications. Hence, producing crop type maps from remote sensing was addressed in many studies, focusing more on mapping herbaceous crop types rather than horticulture crops [9].

Machine Learning (ML) algorithms such as Multilayer Perceptron (MLP), decision tree, maximum likelihood, Linear Disclination Analysis (LDA) and Support Vector Machine (SVM) [16,17] have been used to analyze optical and microwave data for crop analysis, including green cover, pigment, growth stage, crop geometry, equivalent water content, stress conditions, and vegetative indices [18] but suffers from high false positive (FP) and false negative (FN) rate [19]. Therefore, Ensemble Learning (EL) [20] fuses individual classifiers prediction to create accurate predictions result for numerous complicated classification tasks [21]. Despite the computational burden, EL arises as a winner technique in numerous competitions to enhance accuracy [22]. The trick to improve ensemble performance is to select the optimum ensemble approach for loosely correlated classifiers. In general, increasing model complexity reduces mistakes owing to reduced model bias. However, due to the large variation, the model begins to overfit. EL tends to maximize model complexity by balancing bias and variance errors. In short, combining many weakly linked models with various methods yields a more powerful and accurate results considering that the diversity between individual models is the key to a resilient ensemble model.

In this paper, we design a crop identification approach based on the basis of ensemble learning and big data technologies. The proposed approach constructs a pool of base classifiers. Then, a voting schema based on each individual classifier weight was proposed. Apache Spark [23], a standard framework for distributed computing of huge data, is used due to its ease-of-use and greater performance than Apache Hadoop. Various experiments were conducted on El Salheya, Ismaili governate. Datasets of crop classes were integrated with the collected sentinel-2 imageries to improve crop classification results. The proposed approach improves the generalization of crop identification performance at a large scale. The main contribution can be summarized as.

- A scalable Apache Spark solution was employed to efficiently process enormous amounts of data. More Apache Spark nodes improve data processing linearly.
- A weighted voting schema was proposed based on individual classifier performance.
- Experiments were conducted on collected dataset for El Salheya, Ismaili governate. Experimental findings were compared to other classifiers and traditional ensemble techniques.

2. Related work

Recent work had successfully integrated EL to boost the performance of crop classification. In this section, the ensemble learning approaches are briefly discussed, then recent work related to machine learning in crop classification is presented.

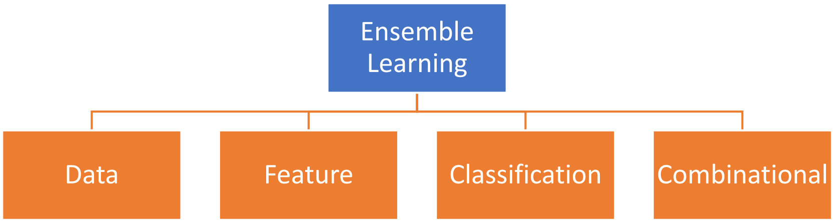


Fig. 1. Ensemble learning taxonomy.

2.1. *Ensemble learning*

Traditional machine learning approaches struggle with large volumes, high-dimensional, noisy data due to their lack of capturing discriminative patterns and features. Ensemble learning [20–22] combines multiple learning algorithms predicting outcomes to improve knowledge discovery and performance. Several ensemble approaches have been introduced in recent decades to meet various application needs [24]. Fig. 1 depicts the four major ensemble learning approaches.

In data ensemble approaches [25], the training data is split into different subsets using resampling methods then these subsets are used to build individual base classifiers. Examples of resampling methods include random selection with or without replacement, leave one out, etc. To fuse the results of all individual classifiers, various voting schema were implemented. These techniques are effective when base classifiers are strongly influenced by small changes in training datasets.

In feature level approaches [26], the training dataset is utilized to extract multiple feature views. Hence, different base classifiers were trained using individual views, then aggregation of the outcomes of these classifiers. Another common practice is to combine different feature views in training base classifiers to generalize better.

In classifier level [27], a heterogeneous or homogeneous classifier pool were trained using training dataset then the outcomes of these classifiers were integrated using rules to minimize the impact of bias and boost the overall performance. These classifies developed either sequentially or concurrently.

In combinational level [27], a pool of heterogeneous or homogenous of classifiers was incorporated with different parameters like injecting randomness into base classifiers. Then, each outcome of the base classifiers in the pool was fused using some rule to boost performance of classification of regression.

2.2. *Machine learning in crop discrimination*

Several ML techniques have been presented in literature for crop mapping and identification. Some of these approaches have performed better than others. The literature shows that the benefits of several algorithms might be merged into an ensemble to increase the performance. Ensemble learning techniques had been widely investigated in several area. Table 1 summarizes numerous contributions, datasets used in remote sensing domain especially crop identification topic.

3. Material and methods

An ensemble approach for crop discrimination was introduced based on big data technology. Fig. 2 depicts an overview of the

Table 1
Summary of crop discrimination using machine learning methods.

Ref.	Approach	Datasets	Outcomes
[28]	CNN, ANN, SVM and RF	Multi-source dataset (Spectroscopy, RGB and HS imageries)	Robotic weed control system
[29]	RF	Sentinel-2A time series	OA (88%), kappa (0.84%)
[30]	SVM, DT, K-NN and ML	Sentinel-2 images	OA (77.2%) with SVM and RF.
[31]	DT	Multi-source dataset (Multi-polarized SAR, Radarsat-2, and Sentinel-2)	OA (66%) using single date Sentinel-2 with 2 date Sentinel-1, OA (89.5%) by incorporating Radarsat-2 data.
[32]	Kernel PCA, and SVM	Radar Sat-2 Images	KPCA-based SVM suppress SVM by 7% in OA incorporates temporal dataset.
[33]	ANN, SVM, RF and K-NN	World View −2	ANN is an efficient to address UAV multispectral data.
[34]	LR, EN, KNN, SVR.	Soil Sampling Data	KNN is the worst performance for potato yield estimation
[35]	PSO was adopted to select the most effective features for ANN, KNN classifier.	Hyperspectral data (EO-1)	The proposed ANN-BA classifier boosts KNN performance in the number of misclassified cases.
[36]	SVM, RF, CART, Sequential Feature Selection approach	Multi source dataset (Sentinel-1A and 2 A)	SVM outperforms the RF and CART using combined optical and SAR datasets.
[37]	RF	Multi-source timeseries dataset (Landsat 7/ETM+, Landsat 8/OL, SPOT 6 and 7, Sentinel-1)	Time series dataset was adopted to highly obtain crop discrimination over the season and red, NIR, and SWIR bands were the most important features.
[38]	SVM and Binary encoding (BE)	EO-1 Hyperion imagery	OA (90.44%).
[17]	PCA, Minimum Noise Fraction (MNF), Wavelet Transform Fisher Linear Discriminant analysis (LDA) and SVM	HS image data collected by imaging spectrometer.	(8 bands) > OA (85%). (15 bands) - > OA (90%).
[39]	ANN and PCA	TERRA/AQUA-Modis and Landsat-OLI	OA (89%)
[40]	SVM and RF	Unmanned Aerial Vehicle (UAV) images	SVM achieved the best crop classification based only on spectral information.
[41]	Maximum Likelihood and Minimum Distance	Spot-5 images	
[42]	Polarimetric Correlation Coefficients.	PolSAR dataset.	*P*- and L-band data effectively discriminate crops while *C*-band data slightly overlapped classification.

proposed approach for four main stages: data collection and preprocessing, feature extraction, classification pool, ensemble schema. First, a time series of multi-source satellite imageries were collected, and their necessary pre-processing was performed. Then, an automated environment was set up to extract physical and handcrafted features. Next, the generated features are stored in a distributed storage system and processed using Apache Spark. We used five diverse classifiers to build a diverse pool and the proposed weighted voting schema is implemented. Finally, the proposed architecture was evaluated on common metrics including overall accuracy, recall, precision, F1-measure and Matthews correlation coefficient (MCC).

3.1. Data collection and preprocessing

El Salheya is located in Ismaili governorate, Egypt. It is positioned between longitudes 32.00449°E and latitudes 30.746557° N. In this work, Sentinel-1 consists of two identical satellites conducting *C*-band SAR imaging at 5.6 GHz (5.4 cm wavelength) with a 12-day revisit duration (6 days considering both satellites).

The Sentinel-2 dataset is Level-1C, which contains ortho-rectification and sub-pixel spatial registration. Sentinel-2 Level-1C consists of 110 km 110 km tiles in UTM/WGS84 projection and offers TOA reflectance. A total of 106 images were collected from Sentinel-2 and Sentinel-1 through the European Space Agency (ESA). In addition, field data was collected to train base classifiers, and validate the results. Samples were extracted during several field campaigns carried out for both summer and winter seasons in the period between Jan. 2019 and Des. 2021 and supplemented by other samples extracted from Sentinel- 1 and 2. Fig. 3 depicts the spatial distribution of the utilized crops in the study area in summer and winter seasons. For summer season, the crop classes are Green Onion, Penaut, Selage, Tree Crops, and uncultivated area. In winter season, the classes are AlfaAlfa, Onion, Potato, Sugar Beet, Wheat, Tree Crops, and Uncultivated. In all, the dataset is balanced as it contains nearly equal numbers for each class. Data preprocessing is mandatory to transform the collected data into a format adequate to the data presentation. The implemented data preparation procedures include data filtering, data labelling, replacing missing numeric values by average value.

3.2. Feature extraction

An automated environment using Python programming was set up to extract physical and handcrafted features, as shown in Fig. 4. The fundamental step in any image-based classification is the feature extraction where the image is transformed into useful information by performing mathematical operations to extract handcrafted features. Various spectral and textural features are anticipated to be effective for crop classification. The features that represent the characteristics of an entity based on reflectance values of the satellite image bands are referred to as spectral features. The spectral features considered in this study are briefly described in Table 2.

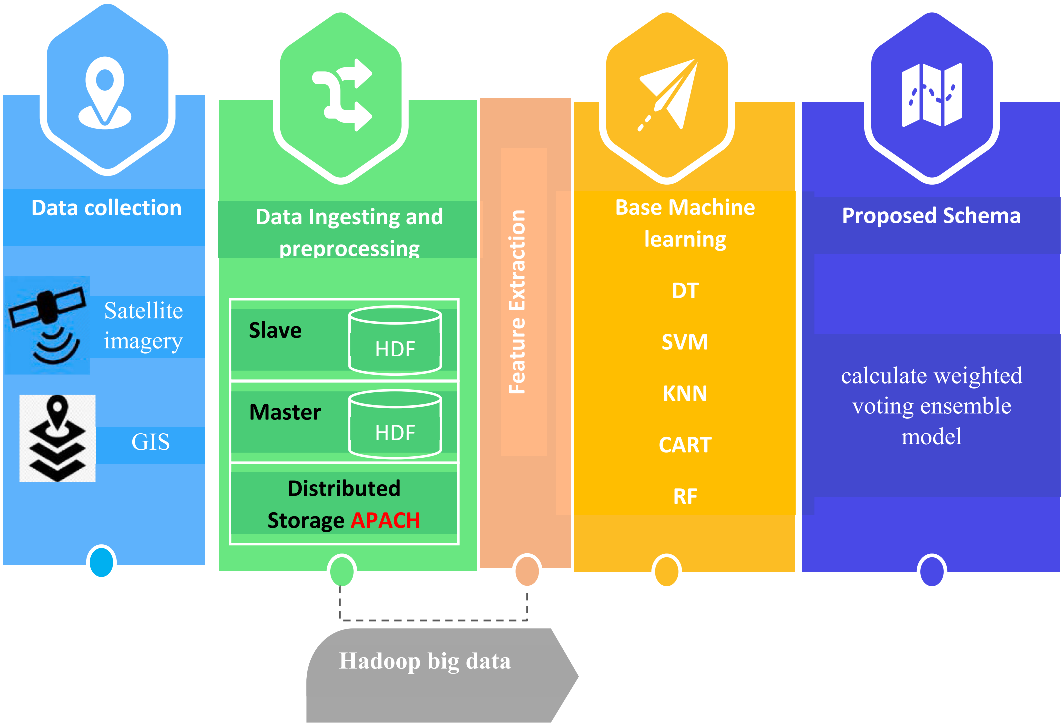


Fig. 2. An overview of the proposed architecture for crop identification using a big data framework.

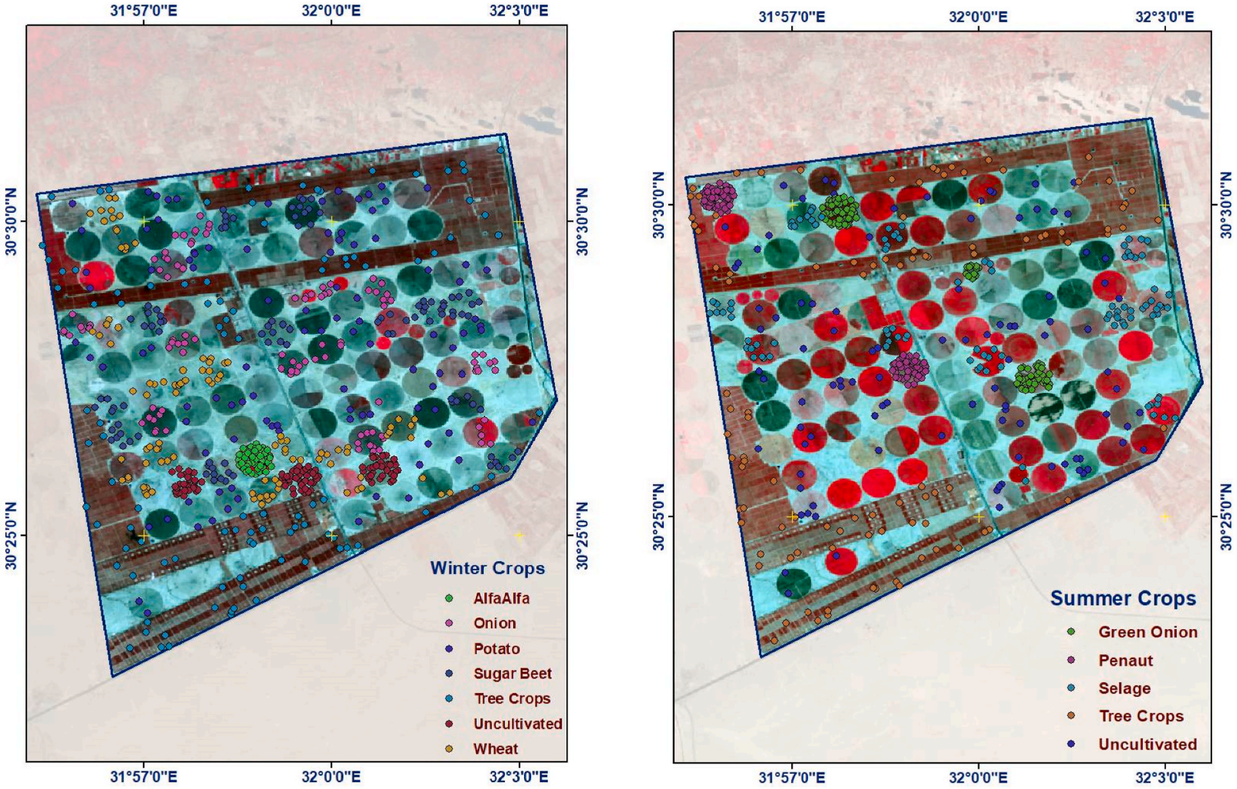


Fig. 3. The spatial distribution of the chosen crop types in summer and winter seasons.

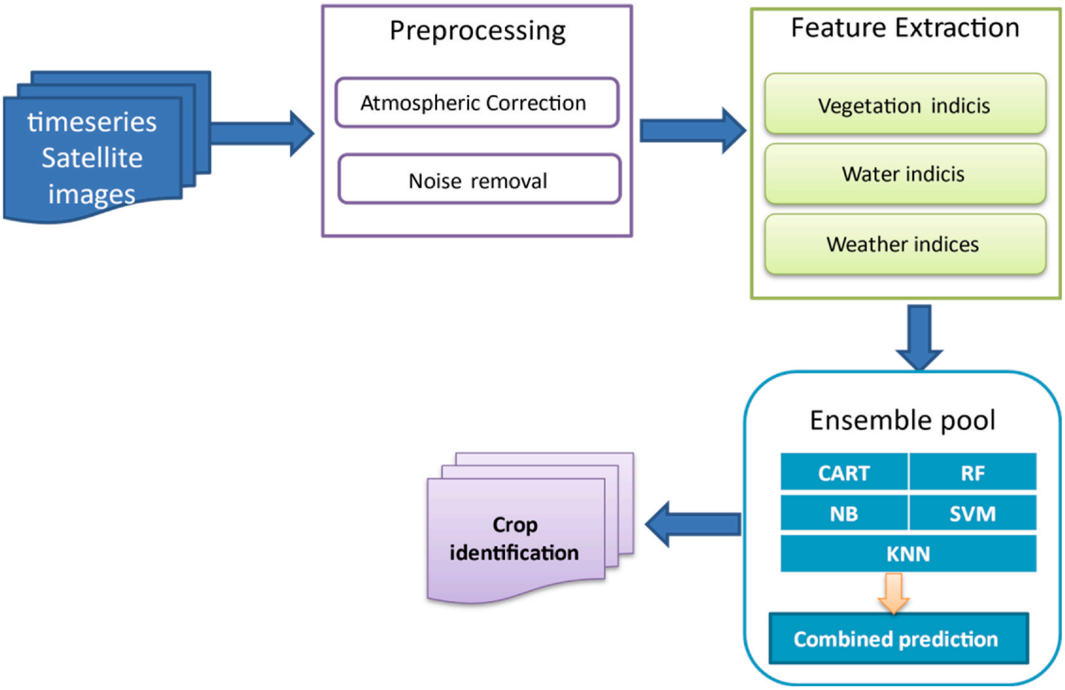


Fig. 4. The proposed ensemble model detailed phases for crop identification.

3.3. Classifiers pool

A diverse pool of classifiers was created to attain high accuracy performance includes Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN).

3.3.1. Decision tree

Decision Trees divided the input data into local regions by a sequence of recursive splits. The tree consists of decision nodes and leaf nodes. Leaf nodes are responsible for the prediction according to the local model associated with that node. In training phase, a split threshold is chosen to reduce mean square error during the recursive split to be less than or equal to an acceptable threshold.

3.3.2. Random forests

The Random Forest model is built on the basis of ensemble learning, to merge several decision trees (the forest) that is built on a randomly divided dataset using a randomization algorithm to generate the final output. The random forest clusters the predictions based on the majority votes received throughout the voting process. The bigger the number of trees in the forest, the better the accuracy, and the less likely to overfit.

3.3.3. Naïve Bayes

The Naive Bayes classifier (NB) is a straightforward yet powerful technique in rapidly developing setting. NB classifier is a probabilistic classifier, which means it makes predictions-based on likelihood. It is dubbed naive because it assumes that the existence of one characteristic is unrelated to the occurrence of other traits. It is termed bayes because it is based on Bayes' Theorem, which is used to assess the likelihood of a hypothesis when previous knowledge is available. It is conditional on the likelihood.

3.3.4. Support Vector Machine

Support Vector Machine (SVM) is a nonlinear supervised classifier characterized by separating hyperplane. The decision hyperplane helps in deciding the boundaries between data points with different labels within the training dataset. SVM was designed to find the optimal hyperplane for the training data. Accordingly, the decision plan helps categorize the new data points.

3.3.5. K-Nearest Neighbors

The K- Nearest Neighbors (KNN) method is one of the simplest and earliest techniques that achieves competitive results especially when combined with prior knowledge. A typical K-NN classifies each unknown occurrence in the training dataset based on the closest K-NN neighbors. The distance measure metric used to determine the closest neighbors influences the performance. In the absence of prior knowledge, most K-NN classifiers employ basic Euclidean metric to measure the distance between training data. Other distance metrics include Minkowski and Chebyshev.

3.4. Proposed ensemble schema

Assume the training dataset to be $\{(\mathrm{x}_1, \mathrm{y}_1), (\mathrm{x}_2, \mathrm{y}_2)\ldots(\mathrm{x}_\mathrm{n}, \mathrm{y}_\mathrm{n})\}$, where $\mathrm{x}\ \varepsilon\ \mathrm{X}^\mathrm{w},$ w is feature vector dimension, n is the number of training samples. A classifier $\Psi : \mathrm{X} \rightarrow\ \Omega$ maps the input feature x into a set of potential class labels set $\Omega\ = \{\omega_1, \ldots..\omega_\mathrm{n}\}$.

A set $\Psi = \{\Psi_{1,} \ldots..\Psi_\mathrm{k}\}$ represents classifiers in diverse pool that could map the input feature set into possible class through a score function. The majority voting mechanism is traditional score functions that counts the base classifier outputs for each class and output the most votes, as described in equation (1):

$$\Psi_{SUM} = \underset{\omega_i}{\mathrm{argmax}} \sum_{\mathrm{k}=1}^{\mathrm{k}} I(\Psi_k(x), \omega_i), \tag{1}$$

where $I(.)$ is the indicator function with the value 1 in the case of the correct classification of the class described by the feature vector x, i.e. when $\Psi_\mathrm{k}(\mathrm{x})\ = \omega_i$. In this context, the output class label was defined by weight each classifier participated in our pool based on its performance and Bayesian weighted voting was considered in voting calculation.

Table 2
The adopted vegetation and water indices.

Indices	Equation
Normalized Difference Vegetation Index (NDVI)	NDVI = (NIR-RED)/(NIR + RED)
Normalized vegetation Index (NVI)	NVI = (COASTAL - BLUE)/(COASTAL + BLUE)
Difference vegetation Index (DVI)	DVI = COASTAL - BLUE
Ratio vegetation Index (RVI)	RVI = COASTAL/BLUE
Enhanced vegetation index (EVI)	EVI = 2.5 * ((NIR - RED)/(NIR + 2.5 * RED +1))
Soil Adjusted vegetation index (SAVI)	SAVI = ((NIR - Red)/(NIR + Red + L)) x (1 + L)
Modified Soil-Adjusted vegetation index (MSAVI)	MSAVI= (2 * NIR + 1 – sqrt ((2 * NIR + 1)2–8 * (NIR - RED)))/2
Normalized Difference Water Index (NDWI)	NDWI = (NIR- SWIR)/(NIR + SWIR)

3.5. Experimental setup

The proposed benefits from open-source component, which include Hadoop Distributed File System (HDFS), to handle the distributing file storage of massive satellite images. Apache Spark is used for efficient preprocessing of big data, while Python is adopted for in-depth analysis of large amounts of data. The experiments were carried out on Apache Spark cluster consisting of a master and two slave nodes, with Ubuntu operating system. The master node has Intel Core TM i-7-550. A brief description of the used open-source software components in our architecture is described in Table 3. Technically, the collected data was split per crop to 60% for training, 20% validation, 20% testing.

3.6. Evaluation metrics

To evaluate our proposed architecture, we employed various evaluation metrics that are commonly used in classification problems as described in Table 4. Typically, True positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN).

4. Results

We conducted many experiments using a big data environment for crops classification. First, five base ML classification models are constructed using 5-fold cross-validation and evaluated using various assessment metrics as given in Table 5.

Fig. 5 compares five base classifiers on the basis of precision, recall, F1-score, accuracy, and MCC, respectively. From Fig. 5 (a, b), the SVM and DT classifiers recorded the highest and lowest precision and f1-score values, respectively. According to Fig. 5 (c, d), the accuracy and F-measure are largest and lowest for the SVM and DT classifiers, respectively. Additionally, SVM provided the highest average accuracy, 85%, followed by 80% for both RF and DT (Fig. 5). According to Fig. 5, SVM achieves the best score using MCC, 84% (representing the greatest correlation between predicted and actual data labels), followed by DT, 76%.

Next, we conducted a comparative analysis between the proposed ensemble and traditional ensemble methods (majority voting, stacking) on the basis of considering all base classifiers. The obtained classification results in terms of precision, F-score, Overall Accuracy (OA) (%), and Matthews correlation coefficient (MCC) for the proposed ensemble, traditional approaches (majority voting, and stacking considering all of classifiers) are illustrated in Table 6.

From Fig. 6, the obtained values for traditional ensemble approaches (majority voting, stacking) and proposed method can be visualized. It clearly reveals that the obtained value for proposed approach is the best. Also, Fig. 7 compares traditional and proposed ensemble approaches. Vividly, the proposed ensemble has the highest precision/F-Measure.

Next, Fig. 7 compares the obtained OA (in percentage) between base classifiers, traditional and proposed ensemble methods. It can be observed that all ensemble methods achieve better accuracy compared with SVM (best base classifier). In addition, the proposed ensemble schema suppresses traditional ensemble methods and attains better accuracy in comparison with majority voting and stacking. It worth to be noting that the same accuracy had been achieved by stacking all classifiers and stacking only top 3 base classifiers. As a result, the use of the best set of classifiers instead of all classifiers decreases calculation time.

Finally, Fig. 8 illustrates the obtained MCC values for the performance of base classifiers, traditional ensemble methods, and the proposed crop identification schema. It can be observed that the value of obtained MCC achieved almost complete mark for each ensemble method which vividly ensures that ensembles achieve high correlation between predicted and actual data labels. The proposed ensemble schema attains the highest MCC value. Thus, the proposed ensemble schema may be utilized to increase the generalization performance for crop identification.

5. Discussion

This study showed the effectiveness of integrating ensemble learning with big data platform to help in large scales crop identification utilizing medium resolution satellite images remote sensing. Moreover, different vegetation indices calculated to boost crop classification in different seasons. The changes in VI's trends of different crops added value to the input data variation for different applied algorithms. According to Ref. [43], the highest classification accuracy achieved of 95% with a voting classifier ensemble in crop mapping in United States using Google earth Engine (GEE). Using cloud platforms, several high-resolution land cover/use maps at the global scale were recently produced [44,45]. However, these datasets have one category for cropland, including all types of herbaceous crops, horticulture crops. The detailed information on the crop types is essential for various agricultural applications. Hence, producing crop type maps from remote sensing data was intensively addressed in earlier studies [46–48], with more focus on the classification of herbaceous crops. In Ref. [9], GEE cloud platform was used for tree crop mapping in Egypt and the comparative

Table 3
Open-source software components.

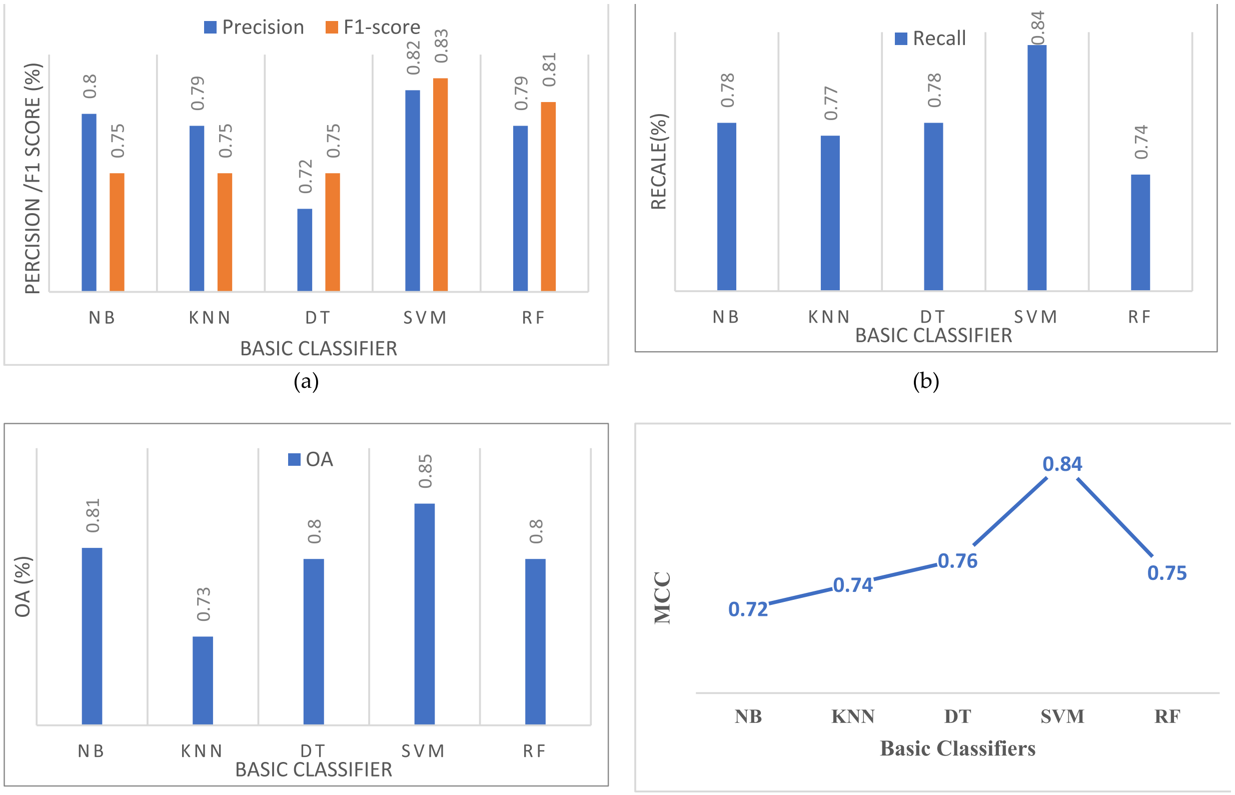
HDFS	Hadoop Distributed File System is commonly used in large scale distributed data due to fault tolerance, fast, and simplicity.
Apache Spark	Apache Spark is a robust and scalable processing engine that utilizes resilient distributed dataset (RDD) due to fault-tolerant units. Compared to Hadoop and MapReduce, it has notable boosting performance.
Python Language	Python is utilized for data-intensive analysis since it has a vast ecosystem of scientific libraries and practically all key ML research publications use Python for implementation. PySpark, a Python API for Apache Spark, and Jupyter Notebook were adopted in the development.

Table 4
Evaluation metrics.

True Positive Rate (TPR)	TP/(TP + FN)
True Negative Rate (TNR)	TN/(TN + FP)
False Positive Rate (FPR)	FP/(TP + FN)
False Negative Rate (FNR)	FN/(TP + FN)
Precision	TP/(TP + FP)
F-Measure	(2 × TP)/(2 × TP + FP + FN)
Accuracy	(TP + TN)/(TP + FP + TN + FN)
Matthews Correlation Coefficient (MCC)	(TP × TN)- (FP × FN) $\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}$

Table 5
The obtained classification results (%) using five base classifiers.

Classifier	Recall	Precision	F1-score	OA	MCC
NB	0.78	0.8	0.75	0.79	0.72
KNN	0.77	0.79	0.75	0.73	0.74
DT	0.78	0.72	0.75	0.8	0.76
SVM	0.84	0.82	0.83	0.85	0.84
RF	0.74	0.79	0.81	0.8	0.75



(a) (b)

(c) (d)

Fig. 5. A comparison of five base classifiers on the basis of a) Precision/F1-score, b) Recall, c) overall accuracy (OA), and d) MCC.

accuracy analysis showed that using time series Sentinel-2 spectral bands (SBs) and vegetation indices (VIs) can classify various tree crop types with very high accuracy (96%), while using Sentinel-1 polarization channels with some added Sentinel-1 textural features could yield a high classification accuracy (85.2%). The results shown high over all accuracy 0.84, 0.85, 0.86, 0.86, 0.85, 0.87, 0.91 and 0.91 for Majority Voting (All classifiers), Majority Voting (SVM, RF, DT), Weighted Voting (All classifiers), Weighted Voting (SVM, RF, DT), Stacking (All classifiers), Stacking (SVM, RF, DT), Proposed Schema (All classifiers) and Proposed Schema (SVM, RF, DT) respectively. The challenge of the proposed approach applicability in Egypt is return to the lack of the agriculture sector data

Table 6
Comparison of traditional ensemble methods with proposed ensemble methods.

Ensemble Method	Precision	Recall	F-score	OA	MCC
Majority Voting (All classifiers)	0.82728	0.84248	0.85569	0.84171	0.84522
Majority Voting (SVM, RF, DT)	0.83029	0.86984	0.88783	0.85462	0.80449
Weighted Voting (All classifiers)	0.81715	0.80944	0.88842	0.85502	0.81693
Weighted Voting (SVM, RF, DT)	0.82138	0.82186	0.86002	0.86352	0.83054
Stacking (All classifiers)	0.81368	0.8509	0.85141	0.8487	0.86826
Stacking (SVM, RF, DT)	0.80616	0.85372	0.86438	0.86984	0.86391
Proposed Schema (All classifiers)	0.89076	0.87007	0.90603	0.9052	0.87593
Proposed Schema (SVM, RF, DT)	**0.89464**	**0.86443**	**0.91788**	**0.91145**	**0.89039**

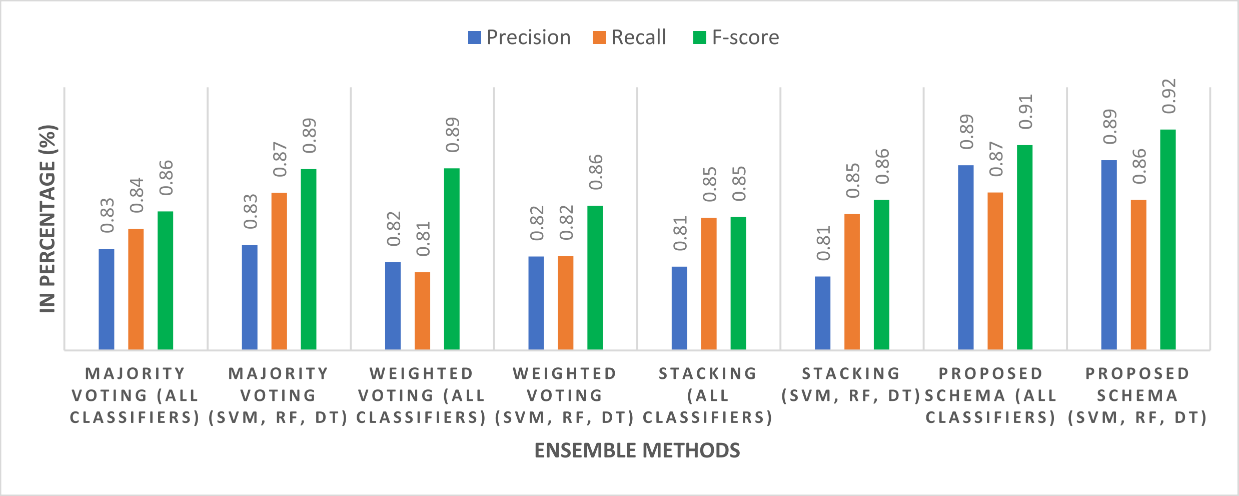


Fig. 6. Comparison of best base classifier, traditional and proposed ensemble methods in terms of precision, recall and F-score.

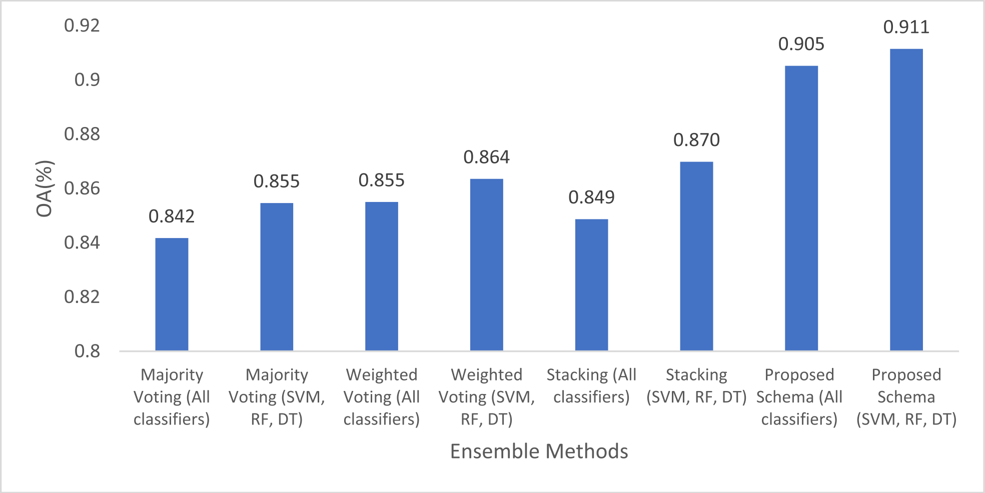


Fig. 7. Comparison of best base classifier, tradition ensemble and proposed ensemble method using accuracy.

infrastructure and fragmentation of agricultural holdings in the majority of old heavy texture soils in Nile delta and Valley.

6. Conclusions

Crop classification plays an important role in smart agriculture technology. Satellite imageries are vital in crop classification due their adequate cost and time. However, the exponential growth of collected imageries emphasizes the need to integrate big data

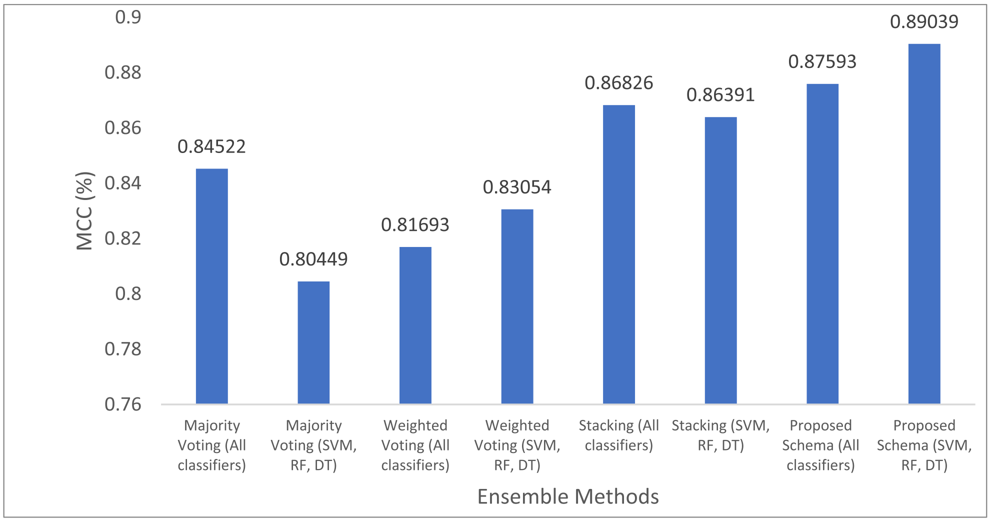


Fig. 8. Comparison of base classifiers, tradition ensemble, and proposed ensemble method using MCC.

technology in agriculture. Traditional crop identification approaches have been proved to be ineffective in big data setting. Therefore, we developed a crop identification approach based on ensemble learning and big data technology. The proposed approach is composed of four stages: data collection and preprocessing, feature extraction, classification pool, ensemble schema. Apache Spark is used as framework for distributed computing. In data collection and processing stage, different preprocessing techniques were applied to the collected imageries. In feature extraction, eight vegetation and water indices were computed. Next, a pool of five base classifiers (DT, RF, NB, SVR, and KNN) was constructed. Finally, the proposed weighted ensemble schema was computed based on the performance of each of the base classifiers. Experiments were conducted on El Salheya, Ismaili governate and results indicate that the proposed approach is superior to other comparative baseline methods in crop identification from satellite imageries. The proposed ensemble schema improved precision, recall, F-score, OA, and MCC by 6.5%, 1.9%, 4.4%, 4.9%, and 4.7%, respectively. In the future, we plan to use the proposed approach to identify other important crops in Egypt. Moreover, we consider integrating the proposed model to other imbalanced crops dentification dataset to further test the effectiveness of the proposed method.

Author contribution statement

Sayed Ahmed, Amira S. Mahmoud, Eslam Farg, Amany M. Mohamed, Marwa S. Moustafa, Ahmed M. Saleh, Mohamed A.E. AbdelRahman, Hisham M. AbdelSalam, and Sayed M. Arafat: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by Science, Technology& Innovation Funding Authority (STDF), Egypt under grant (ESIP 2019) project ID (33,547).

Data availability statement

Data included in article/supp. Material/referenced in article.

Declaration of interest's statement

No conflict of interest.

References

[1] R. Springborg, Egypt's economic transition: challenges and prospects, International Development Policy| Revue internationale de politique de développement 7 (2017).
[2] E. Collado, A. Fossatti, Y. Saez, Smart farming: a potential solution towards a modern and sustainable agriculture in Panama, AIMS Agriculture and Food 4 (2) (2018) 266–284.
[3] V. Saiz-Rubio, F. Rovira-Mas, From smart farming towards agriculture 5.0: a review on crop data management, Agronomy 10 (2) (2020) 207.
[4] J.A. Delgado, N.M. Short Jr., D.P. Roberts, B. Vandenberg, Big data analysis for sustainable agriculture on a geospatial cloud framework, Front. Sustain. Food Syst. 3 (2019) 54.

[5] S. Mayr, C. Kuenzer, U. Gessner, I. Klein, M. Rutzinger, Validation of earth observation time-series: a review for large-area and temporally dense land surface products, Rem. Sens. 11 (22) (2019) 2616.

[6] J. Soria-Ruiz, Y. Fernandez-Ordonez, H. McNairn, P.H. Pei-Gee, Corn monitoring and crop yield using optical and microwave remote sensing, Geoscience and Remote Sensing 598 (2009).

[7] H. Huang, et al., Modelling and validation of combined active and passive microwave remote sensing of agricultural vegetation at L-band, Progress In Electromagnetics Research B 78 (2017) 91–124.

[8] E. Farg, S. Arafat, M. Abd El-Wahed, A. El-Gindy, Estimation of evapotranspiration ETc and crop coefficient Kc of wheat, in south Nile Delta of Egypt using integrated FAO-56 approach and remote sensing data, The Egyptian Journal of Remote Sensing and Space Science 15 (1) (2012) 83–89.

[9] M. Nabil, E. Farg, S.M. Arafat, M. Aboelghar, N.M. Afify, M.M. Elsharkawy, Tree-fruits crop type mapping from Sentinel-1 and Sentinel-2 data integration in Egypt's New Delta project, Remote Sens. Appl.: Society and Environment (2022), 100776.

[10] E. Farg, M.N. Ramadan, S.M. Arafat, Classification of some strategic crops in Egypt using multi remotely sensing sensors and time series analysis, The Egyptian Journal of Remote Sensing and Space Science 22 (3) (2019) 263–270.

[11] I.H. Woodhouse, Introduction to Microwave Remote Sensing, CRC press, 2017.

[12] S.R. Sultana, et al., Normalized difference vegetation index as a tool for wheat yield estimation: a case study from Faisalabad, Pakistan, Sci. World J. (2014) 2014.

[13] M. Hosseini, H. McNairn, A. Merzouki, A. Pacheco, Estimation of Leaf Area Index (LAI) in corn and soybeans using multi-polarization C-and L-band radar data, Rem. Sens. Environ. 170 (2015) 77–89.

[14] G.N. da Piedade, L.V. Vieira, A.R. dos Santos, D.J. Amorim, M.D. Zanotto, M.M. Sartori, Principal component analysis for identification of superior Castor bean hybrids, J. Agric. Sci. 11 (9) (2019).

[15] A. Begue, et al., Remote sensing and cropping practices: a review, Rem. Sens. 10 (1) (2018) 99.

[16] I. Nitze, U. Schulthess, H. Asche, Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification, in: Proceedings of the 4th GEOBIA, Rio de Janeiro, Brazil 79, 2012, p. 3540.

[17] B. Liu, R. Li, H. Li, G. You, S. Yan, Q. Tong, Crop/Weed discrimination using a field imaging spectrometer system, Sensors 19 (23) (2019) 5154.

[18] J. Xue, B. Su, Significant remote sensing vegetation indices: a review of developments and applications, J. Sens. 2017 (2017) 135369. https://doi.org/10.1155/2017/1353691.

[19] S. Feng, J. Zhao, T. Liu, H. Zhang, Z. Zhang, X. Guo, Crop type identification and mapping using machine learning algorithms and sentinel-2 time series data, IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 12 (9) (2019) 3295–3306.

[20] J. Rocca, Ensemble methods: bagging, boosting and stacking, Data Sci. 5 (2019).

[21] L. Yang, Classifiers selection for ensemble learning based on accuracy and diversity, Procedia Eng. 15 (2011) 4266–4270.

[22] V. Smolyakov, Ensemble Learning to Improve Machine Learning Results, Stats & Bots, 2017.

[23] A. Spark, Apache spark, Retrieved January 17 (1) (2018).

[24] P. Pintelas, I.E. Livieris, Special Issue on Ensemble Learning and Applications, vol. 13, Multidisciplinary Digital Publishing Institute, 2020, p. 140.

[25] Z. Chen, J. Duan, L. Kang, G. Qiu, A hybrid data-level ensemble to enable learning from highly imbalanced dataset, Inf. Sci. 554 (2021) 157–176.

[26] A. AlSuwaidi, B. Grieve, H. Yin, Feature-ensemble-based novelty detection for analyzing plant hyperspectral datasets, IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 11 (4) (2018) 1041–1055.

[27] D.C. Corrales, A.F. Casas, A. Ledezma, J.C. Corrales, Two-level classifier ensembles for coffee rust estimation in Colombian crops, Int. J. Agric. Environ. Inf. Syst. 7 (3) (2016) 41–59.

[28] W.-H. Su, Advanced Machine Learning in Point Spectroscopy, RGB-and hyperspectral-imaging for automatic discriminations of crops and weeds: a review, Smart Cities 3 (3) (2020) 767–792.

[29] A. Htitiou, A. Boudhar, Y. Lebrini, R. Hadria, H. Lionboui, T. Benabdelouahab, A comparative analysis of different phenological information retrieved from Sentinel-2 time series images to improve crop classification: a machine learning approach, Geocarto Int. (2020) 1–24.

[30] M.G. Maponya, A. Van Niekerk, Z.E. Mashimbye, Pre-harvest classification of crop types using a Sentinel-2 time-series and machine learning, Comput. Electron. Agric. 169 (2020), 105164.

[31] S. Shanmugapriya, D. Haldar, A. Danodia, Optimal datasets suitability for pearl millet (Bajra) discrimination using multiparametric SAR data, Geocarto Int. 35 (16) (2020) 1814–1831.

[32] D. Mandal, V. Kumar, Y.S. Rao, An assessment of temporal RADARSAT-2 SAR data for crop classification using KPCA based support vector machine, Geocarto Int. (2020) 1–13.

[33] S.S. Virnodkar, V.K. Pachghare, V. Patil, S.K. Jha, Remote sensing and machine learning for crop water stress determination in various crops: a critical review, Precis. Agric. 21 (5) (2020) 1121–1155.

[34] F. Abbas, H. Afzaal, A.A. Farooque, S. Tang, Crop yield prediction through proximal sensing and machine learning algorithms, Agronomy 10 (7) (2020) 1046.

[35] M. Dadashzadeh, et al., Weed classification for site-specific weed management using an automated stereo computer-vision machine-learning system in rice fields, Plants 9 (5) (2020) 559.

[36] S. Mustak, G. Uday, B. Ramesh, B. Praveen, Evaluation of the performance of SAR and SAR-optical fused dataset for crop discrimination, Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 42 (3) (2019) 563–571.

[37] C. Lira Melo de Oliveira Santos, et al., Classification of crops, pastures, and tree plantations along the season with multi-sensor image time series in a subtropical agricultural region, Rem. Sens. 11 (3) (2019) 334.

[38] R.R. Surase, et al., Assessment of EO-1 Hyperion imagery for crop discrimination using spectral analysis, in: Microelectronics, Electromagnetics and Telecommunications, Springer, 2019, pp. 505–515.

[39] C.C. Junior, et al., Artificial neural networks and data mining techniques for summer crop discrimination: a new approach, Can. J. Rem. Sens. 45 (1) (2019) 16–25.

[40] G.-H. Kwak, N.-W. Park, Impact of texture information on crop classification with machine learning and UAV images, Appl. Sci. 9 (4) (2019) 643.

[41] J. Soria-Ruiz, Y.M. Fernandez-Ordonez, Crop discrimination using remote sensing data in a region of high marginalization, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2017, pp. 3031–3034.

[42] S.-W. Chen, Y.-Z. Li, X.-S. Wang, Crop discrimination based on polarimetric correlation coefficients optimization for PolSAR data, Int. J. Rem. Sens. 36 (16) (2015) 4233–4249.

[43] S.D. Suchi, A. Menon, A. Malik, J. Hu, J. Gao, Crop identification based on remote sensing data using machine learning approaches for fresno county, California, in: 2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService), IEEE, 2021, pp. 115–124.

[44] K. Karra, C. Kontgis, Z. Statman-Weil, J.C. Mazzariello, M. Mathis, S.P. Brumby, Global land use/land cover with Sentinel 2 and deep learning, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE, 2021, pp. 4704–4707.

[45] R. d'Andrimont, A. Verhegghen, G. Lemoine, P. Kempeneers, M. Meroni, M. Van der Velde, From parcel to continental scale–A first European crop type map based on Sentinel-1 and LUCAS Copernicus in-situ observations, Rem. Sens. Environ. 266 (2021), 112708.

[46] S. Ofori-Ampofo, C. Pelletier, S. Lang, Crop type mapping from optical and radar time series using attention-based deep learning, Rem. Sens. 13 (22) (2021) 4668.

[47] N. You, et al., The 10-m crop type maps in Northeast China during 2017–2019, Sci. Data 8 (1) (2021) 1–11.

[48] Q. Pan, M. Gao, P. Wu, J. Yan, M.A.E. AbdelRahman, Image Classification of Wheat Rust Based on Ensemble Learning, Sensors 22 (2022) 6047. https://doi.org/10.3390/s22166047.