The Impact of Data Analytics in Digital Agriculture: A Review

Nabila Chergui
Faculty of Technology, Ferhat Abbas University, Setif 1
MISC Laboratory Abdel Hamid Mehri University
Constantine 2, Algeria
nabila.chergui@umc.edu.dz

M-Tahar Kechadi
School of Computer Science
University College Dublin
Dublin, Ireland
tahar.kechadi@ucd.ie

Michael McDonnell
School of Business
University College Dublin
Dublin, Ireland
michael@ucd.ie

***Abstract*—The advanced development in Information and Communication Technologies (ICT) and its adoption in the agriculture area open the field to the appearance of Digital Agriculture; which created new processes for making farming more productive, efficient, controllable while respecting the environment. Data science (and machine learning) is among key of information technology used in Digital Agriculture for their ability to analyse a vast amount of data to extract new knowledge and to help agriculture understand better the farming tasks and make better decisions. Big data in its turn offers a support to farmers to extract new insights from their data and to make more accurate decision. This work presents a systematic review of methods and techniques of (data & big data) mining and their applications to Digital Agriculture from the big data view point. In this study, we will focus on crop yield. We first introduce the crop yield management process and its components, and then we focus on the crop yield monitoring. We then present a classification of data mining techniques applied for the crop yield monitoring tasks. This is followed by discussing each category of the classification throughout a panoply of existing works and show their used techniques, then we provided a general discussion on the applicability of big data analytics into the field of digital agriculture.**

***Index Terms*—Digital Agriculture, Data Analytics, Crop Yield Monitoring, Data Mining, Big Data, Machine Learning.**

I. Introduction

The increasing demand for improving the productivity of both small and large farms by reducing resource costs, such as water, fertilisers and pesticides, requires the use of new advanced farming and management techniques. On the other hand, the improvement of productivity does not have to be in the cost of decreasing the quality of products which will harm the health or create damages to the environment, because of the excess use of fertilisers and pesticides and other agricultural inputs. Digital Agriculture (DA), (or digital/smart farming) [1]–[3], is an advanced approach that makes farms and the act of farming smarter by integrating the use of digital and smart tools like (sensors, cameras, satellite, drones, GPS, etc.) in conjunction with Artificial Intelligence (AI) techniques, Internet of Things (IoT), data mining and data analytics to improve the agricultural practices, to enhance the productivity and to optimise the use of resources by providing insights and decision-making supports to farmers. DA can for example controls a crop nutrition by finding the optimum fertilisation program for each field, its optimum irrigation program, and can help farmers to react differently for each part of the field.

The DA involves the development, adoption and iteration of all above-mentioned technologies in the agricultural sector in different spatial contexts [4].

DA is data-driven solutions, which across the use of ICT, AI and Data analytics, permits to farmers to adopt these solutions for their businesses. These solutions can vary and cover multiple activities of farming, including assessing risks and disasters, producing predictive models and so on. We can summarise the benefits of DA to agriculture in:

- It provides the farmer with useful information and supports their decision making with regards to how much, when and where to apply nutrients, water, seeds, fertilisers, and other chemicals and agricultural inputs.
- It sustains the environment [5] and it helps on providing healthy products, since it allows to vary the amount of input resources (irrigation, fertilisers and pesticides) and even seeds used for crop production, and applying those inputs with exact quantities in each field
- Data-driven enable farmers to access to sophisticated management solutions against climate change and other environmental challenges and natural events. through these solutions, farmers can continuously monitor crop health and can obtain on time alert to likely problems with pests or disease or even climate change.
- From the marketing view, farmers can also benefit from advanced models that give insights on the market and which products could bring more profits to them.

In the past, the full potential of DA was not possible. Nowadays, data gathering and data mining & analytics techniques are commonplace in every sector of the world economy. This was made possible with technological advances (sensor devices, satellite images, advanced weather stations, etc.) and also advances in digital devices and the ability to store and process nearly everything. Today we can store vast amounts of data. Besides, with the use of advanced data mining techniques, we can extract novel and useful knowledge from these large volumes of historical datasets, which can help us

978-1-7281-6403-8/20/$31.00

understand the behaviour of both crops and farmed fields and, therefore, use efficient management techniques of the whole farming industry, such as field and wasteland management, crop and pest management, soil classification, etc.

Crop management is a key task in DA, as it impacts directly on the crop production; it assists crops from soil preparation and seed selection, watering, and so on, until the harvest day, and can go beyond post-harvest. It offers the yield variability that exists in farm fields, permitting producers to figure out how management skills and environmental factors affect crop production [5]. This results on offering valuable feedback to farmers enabling them to take better decisions [6] at real-time and monitor the farm proactively. The crop management process, if we ignore the marketing of products, can be devised into five sub-processes as described below, and as presented in in Figure 1.

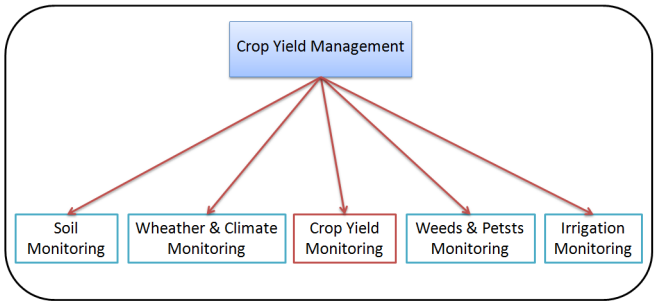

Fig. 1. The crop yield management components.

- **Soil monitoring:** it aims at studying the nature of the soil to select the type of culture to plant. Also, it controls the concentration of fertiliser and, hence, the quantities to use depending on the soil type, and it controls the irrigation operation based on the soil moisturising.
- **Weather & climate monitoring:** This deals with the monitoring of temperature, wind and other environmental factors that can affect the crop.
- **Weed and Pest monitoring:** it controls the insects and weeds that can affect each type of crop, and the type and quantities of pesticides and herbicides that should be applied at a given location of the farm;
- **Crop Yield monitoring:** it aims at monitoring crop conditions during the growing season, the estimation of the crop yield, crop quality, detection and protection of crops from diseases, the delineation of management zones of yields, and all the other related crop operations.
- **Irrigation monitoring:** it takes into consideration the amount of water needed by each type of crop and the irrigation rate, which also depends on weather conditions, the season, the soil type, and the crop's growth cycle.

Vast amounts of data were gathered from each of these processes. Its exploitation using the potential power of data analytic techniques will offer a solid decision-making support to farmers and will have a significant impact on crop production and on the environment conservation. Data mining has several applications in crop yield management, such as the understanding of vegetation variables, soil mapping and classification, zone management, weather forecasting, disease protection, prediction of the market crop prices, rainfall forecasting, weed detection and yield prediction, etc. In this work we will focus on the crop yield monitoring, which is part of the management process.

It is clear, and mainly in DA, that more data we collect more insights we can acquire from it and more accurate the results will be. Nowadays, collecting data is not a hurdle anymore. We live of the era of big data and IoT, and very large amounts of data can be collected from various sources. For instance, data can be originated from crop yield, patterns and rotations, weather, climate and environmental conditions and parameters, soil types, moisturising and nutrients, and from farmers' records on yields and other factors. This collected data is not only big, but also heterogeneous in types and quality. Therefore, its analysis is very challenging.

Part of this data heterogeneity in DA came from the way the data were collected, as each data collection technique has different characteristics in its accuracy, validity, and impact on farmland.

Accordingly, another part of this heterogeneity caused by the type of the used devices to collect data, different sensors, different records and different cameras, etc.

Considering these facts, and in light of its source and nature, data can belong to one of the following type classes: historical data, sensor data, image data or satellite data.

The remainder of the paper is organised as following: Section II presents related works on the application of data mining & analytics, AI and machine learning to crop monitoring. In addition it describes a classification of data mining techniques applied for crop yield monitoring. Section III discusses classification techniques for crop yield. Section IV shows the techniques used for the prediction of crop yield for both types of data. Section V exhibits the techniques used to protect crops from diseases, pests and weeds. Section VI presents techniques used to detect crops and estimate yields. Section VII is dedicated to clustering techniques used for crop yield. The evaluation of several existing works is presented in VIII. Finally, we conclude the work in Section IX.

II. Crop Yield Management

Various surveys have been addressed the application of data (mining & analytics) to crop yield management. [7] Discussed the yield forecasting using machine learning and integrating agrarian factors. [8] provided a systematic review on the use of computer vision and AI in DA for grain crops.

[9] reviewed the utilisation of big data analysis in agriculture. The authors concluded that the applicability is still at its early stage and several barriers need to be overcome despite the availability of the data and tools to analyse it.

[10] presented a review of advanced machine learning methods used for the biotic stress detection in crop.

An early similar study was presented in [11], where the authors studied four very popular learning approaches in the area of agriculture; Artificial Neural Network (ANN),

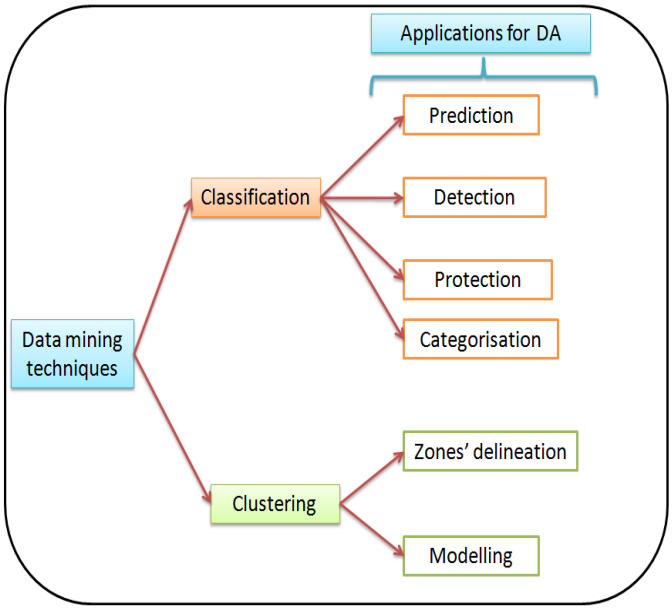

Fig. 2. Data mining techniques applied for crop yield monitoring.

K-means, K-Nearest Neighbour (KNN) and Support Vector Machine (SVM).

The study presented in this paper is not just an update about what has been done in the previous surveys. As big data is commonplace nowadays, the first objective is to examine the application of big data analytics to DA and more specifically on crop yield monitoring. The second objective is to discuss how it is applied and what are the encountered challenges.

So that, the motivation behind the preparation of this review is to figure out how much the big data is employed in DA, and whether the application of data mining techniques in DA implies the use of big data too. Besides, if applicable, how big data and data analytics are leveraged to the benefice of DA.

Therefore, the contribution of this study is to present a deeper analysis of the encountered problems in agriculture and resolved by DA, and to provide an overview that focuses on an important and particular problem, the crop monitoring, compared to the above-mentioned surveys. Furthermore, our study highlights the type of data used, the methods and techniques employed and for which class of data mining techniques.

From the analysed literature, the application of data (mining/analytics) techniques to crop yield monitoring is almost limited to classification and clustering. Figure 2 proposes a classification of data mining techniques applied to crop yield.

Based on the type of data, data mining process can use various pre-processing techniques before starting the analysis of data. For instance, for image-based data, we can use the process described in Figure 3. Once data have been cleaned and pre-processed, the resulting data should be of high quality and ready for the analysis. Depending on the question to be answered and whether the historical data was annotated or not, we choose the category of the analysis techniques (classification, clustering, etc.). Classification techniques, for

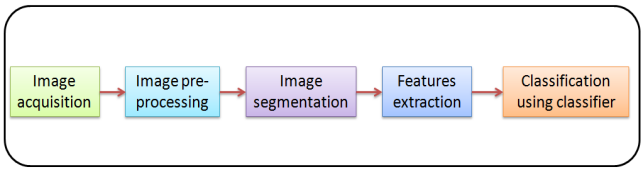

Fig. 3. Image processing approach.

example, are used for the purposes of prediction, detection, protection, and categorisation tasks, which are the most important tasks involved in the crop monitoring process. The following sections discuss the research that has been conducted in each of the four principal tasks of the crop monitoring process.

III. Crop Classification

Many mining approaches have been used based on both collected datasets and the target objective [12]–[17]. In this section we review some of these works.

To identify and classify potato plants and three types of weeds, [12] used a machine vision system, which consists of two subsystems: a video processing subsystem that is able to detect from each frame the green plants; and a hybrid approach that combines ANN and particle swarm optimisation algorithm (PSO) to classify weeds from potato plants. The PSO is used to optimise the ANN's parameters and the ANN for classification. The hybrid approach was compared to Bayesian classifier (BC). The experimental results show that ANN-PSO outperforme the BC.

[13] used a multilevel deep learning architecture to classify land-cover and crop based on multi-temporal multi-source satellite imagery data. They pre-processed the data by segmenting the imagery data and restoring the missing data due to clouds and shadows before classifying it. The proposed hybrid approach was compared to Random Forest (RF) and Multi-layer Perceptron (MLP). They showed that their approach outperforms both RF and MLP and obtained better discrimination for some crops like maize and soybeans.

Deep learning approach has been also used for plants species and weeds classification, based on coloured images issued from six different data sources [14]. They used Convolution Neural Network (CNN) on a dataset consisting of $10,413$ images with 22 weeds and crop species. The CNN model achieved an overall accuracy of 86.2%.

[15] developed a hybrid crop (corn, soybean, cotton, rice) classifier based on satellite images. The approach is an ensemble learner that consists of an ANN, SVM and decision tree (DT), and a combiner to generate a prospective decision. The overall approach generates a recommendation based on both the available expert knowledge and the learners outputs.

An SVM-based classifier for distinguishing crops from weeds based on digital images was suggested in [18]. It achieved an accuracy of 97%.

IV. Crop Yield Prediction

The estimation of crop yield aims to study factors that influence and affect the crop production, such as climate and weather, irrigation practices, natural soil fertility and its physical structure and topography, crop stress, incidence of pests and diseases,etc. It enables efficient planning of resources; an early and accurate prediction of yields can provide a solid base to decision makers to determine if there will be a shortage or a surplus, thus, to react properly according the the situation. In the following we discuss some prediction techniques and showing type of the datasets that were used for the prediction operation.

A. Soil and Weather Data

Crop yield prediction using historical data and time series has been studied for many years, an it is considered among the classical applications of data mining. In 1994, [19] employed a fuzzy logic expert system to forecast corn yield. The model obtained promising results. A year later, [20] used a feed-forward Back-Propagation NN (BPNN) to predict soybean and corn yields. The dataset used was based on soil properties, such as PH, phosphorus, potassium and magnesium saturation's, organic matter, topsoil depth. Other factors, such as weather were not considered. The NN showed promised results as a support of yield variability understanding, but its accuracy is not reasonably good. To improve its accuracy, [21] proposed a projection pursuit regression (PPR) and Multiple Linear Regression (MLR) all with several types of supervised feed-forward NN methods for yield estimation to study the relationships between yield, topographic characteristics and soil properties. The authors added a second phase of experiments to include climatological data. The NN techniques outperformed both MLR and PPR and provided minimal prediction errors. Besides, the results demonstrated that a important over-fitting had happened, in addition that this type of analysis required a larger number of climatological site-years.

Moreover, to analyse the weather aberrations impacting on rice production in mountainous region of Fujian province of China, [22] used a NN model. The historical dataset was collected from 16 locations throughout the region. The weather variables include: daily solar radiation and sunshine hours, daily daily wind speed and daily temperature sum, in addition to the seven different soil types for each location. It was shown that the model is more effective compared to a multiple linear regression model.

MLP were employed for winter wheat prediction by [23], then two different neural networks are considered in [24], and compared with four regression models for yield prediction based on agricultural data yield obtained from a farm in Germany. Networks with MLP and RBF, The Support Vector Regression (SVR) and decision regression tree were implemented. The work demonstrated that the SVR algorithm was the most suitable for this kind of problem.

[25] Demonstrated the applicability of RF to predict the mango crop yield and studied the influence of water supply on yield using four different irrigation regimes. A set of four RF models using a combination of 10 days rainfalls and irrigation data with different input variables including rainfall, irrigation, rainfall and irrigation, and the total water supply model, were developed to estimate the minimum, mean and maximum values for each of the mango fruit yields.

[26] Proposed to use machine learning for pre-season agriculture yield estimation. It employed a Recurrent Neural Network (RNN) fostered by data from multiple resources: precipitation data derived from satellite, soil properties data sets, seasonal climate forecasting and historically observed soybean yield to produce a pre-season forecasting of soybean/maize yield for Brazil and USA.

[27] Conducted a study to compare the predictive accuracy of several machine learning methods (MLR, M5-Prime regression trees, MLP, SVR and K-nearest neighbour) for yield estimation in ten crop data sets. It used data on solar radiation, rainfall, temperature, season-duration cultivar, planting area, etc. It concluded that the M5-Prime is the most suitable algorithm to predict yield for massive crops in agricultural planning.

[28] Evaluated the ability of RF in comparaison with MLR to estimate crop yield (potato, maize, wheat) regarding climate and biophysical variables at global and regional scales. It employed crop yield data from different sources and regions in the USA over 30 years for model training and testing. The results demonstrated that RF outperformed MLR and was capable of predicting crop yields.

The Extreme Learning Machine (ELM) has been employed by [29] to the estimation of the Robusta coffee yield according to the soil fertility properties. The performance of 18 different ELM-based models was evaluated with several combinations of the predictor variables based on the soil organic matter. The MLR and RF have been chosen for the comparison, the results indicated that the EML outperformed the MLR and RF models.

B. Image-based Datasets

[30] Built predictive models for barley, canola and wheat crops from different seasons: pre-sowing, mid-season and late-season, and it explored the worth of gathering data over multiple sources and years into one data-set together with machine learning approaches on the prediction quality. The collected data consisted of yield data, the electromagnetic induction survey EM and gamma radiometric survey, Normalised Difference Vegetation Index (NDVI) and rainfall. RF models were used to predict crops yields using the space-time cube. Performances of the models were improved as the season progressed, since additional information about within-season data were obtained.

[31] Built a model for citrus yield forecasting from airborne hyper-spectral images using ANN. It used an airborne imaging spectrometer to acquire images over a citrus orchard. The work concluded that its obtained results demonstrated that the ANN performed well for the observed hyper-spectral data, and suggested to adopt the use of airborne hyper-spectral remote sensing to predict citrus yield.

[32] Proposed a prediction model for within-field variation in wheat yield using several neural networks types and sensing techniques, based on on-line multi-layer soil data, and satellite imagery for crop growth characteristics. They have chosen the Counter Propagation Neural Network (CP-ANN), XY-fused Networks (XY-Fs) and Supervised Kohonen (SKN) as machine learning approches to test performances of the predicting wheat yield model. Results demonstrated that the SKN model had the best overall performance.

[33] Established a cabbage yield forecasting model using the Green-seeker hand-held optical sensor and the NDVI. The model also sought to measure the application rate of four chemical nitrogen and their relation with the cabbage yield. To evaluate the relationship between NDVI measurements and yield, modified exponential, linear and quadratic functions were employed. By comparison, the exponential equation performed better than the linear and quadratic functions.

[34] Developed two ANN models for early yield prediction of 'Gala' apple trees. These models were analysed 50 RGB images of the trees to identify the fruit. These two models were proved to predict yield accurately.

[35] Extended the model of [36] to estimate the yield for soybean crops, and transfer the learning from Argentina to Brazil using deep learning techniques (Long/Short Term Memory Network(LSTM)), and the Moderate Resolution Imaging Spectro-radiometer (MODIS) satellite imagery as data. To transfer learning, it first initialised the LSTM model with the parameters from a network and trained it on Argentine soybean yields. Then, it removed the last dense layer of the trained model on Argentine data-set and replace it with an untrained dense layer of the same dimensions before training the modified model on Brazilian data-set. The results on both data-sets demonstrated that this approach can learn with success effective features from raw data.

[37] Used the MODIS EVI (Enhanced Vegitation Index) product and ground temperature measurements to estimate corn yield using SVM and DNN. The DNN was able to provide accurate predictions.

V. Crop and Plant Protection

Forewarning, protection and detection of crop diseases correctly and timely when they first appear is a very important task of crop monitoring, it will reduce yield losses and inform and prevent farmers to take effective preventive actions. For detecting crop and plants diseases, several works have used image processing, consequently image-based data and classification techniques for detecting crops diseases' [38]–[42]. The general approaches of these works are almost similar, as described in the process presented in Figure 3. The approaches started by capturing and collecting images for disease plants using cameras, scanners or other sensors. After that, segmentation of plant disease spots, followed by an extraction of features like colour, shape or texture. In the end, the employment of classification methods, such as ANN, BC method, KNN, SVM, to classify disease images.

[38] and similarly [43] have proposed a deep CNN for plant disease recognition and detection from images data-sets containing images of diseased and healthy plant leaves. The former achieved a good performances, and the latter achieved an average accuracy of 96.3%.

[39] Introduced an automatic detection and classification method for crop disease using plant leaf images. This method composed of four phases; it started by the image acquisition, then a pre-processing by creating of a colour transformation structure for the RGB leaf image followed by the application of device-dependent colour space transformation for the colour transformation structure. After that, segmentation of images performed by K-means clustering technique to calculate the texture features. After that, a classification of the extracted features using ANN.

For accurate and early detection of rice crop disease; [40] presented an application of SVM for rice diseases detection's. The accuracy of the detection and classification of these disease spots was very good.

[44] Applied an ANN to discriminate the level of rice panicles infections by fungal. It used hyper-spectral reflectance of rice panicles which was measured through the wavelength with a portable spectro-radiometer in the laboratory after hand cutting of samples from three different fields in China.After that, a LVQ neural network has been used to classify fungal infection levels into one class among (healthy, light, moderate, serious). The obtained results showed a good accuracy.

[45] Investigated the performance of four classification algorithms applied to the problem of classification of Egyptian rice diseases using historical data. In this work, a comprehensive comparative study of four classification algorithms and their performances has been examined, namely: J-48 DT, random trees (RT), Naive Bayes net, and RF. The experimental results indicated that the J-48 DT gave the best results, where the Naive Bayes was the worst.

Deep learning has been used also by [46]–[50].

Among those, [46] used leaves images of healthy and diseased plants and CNN model to carry out the plant disease detection and diagnosis model. It employed a database incorporated 87848 images containing 25 different plants(including unhealthy and healthy plants) in a set of 58 distinct classes of (plant, disease) combinations. It tried different architectures for training in order to get the best performance.

Another work [48] is proposed to identify tomato diseases based on database of 18149 images, using a combination of super-resolution (CNN) and conventional images to improve the spatial resolution of diseased images, and to recover detailed appearances. A super-resolution CNN outperformed the other conventional disease classification methods.

Five different architectures of deep CNNs have been evaluated by [49] for the classification of plant diseases based on images from the open data-set PlantVillage. The evaluated architectures including: DenseNets with 121 layers, Inception V4, ResNet with 50, 101 and 152 layers and VGG 16. The results showed that the accuracy of DenseNets model was

much better than the other models, and that it had ability to improve its accuracy with the growing number of epochs.

[51] Used SVM algorithm which takes NDVI and raw data inputs to develop weed (silverbeet) and crop (corn) discrimination. A conventional plant discrimination algorithm has been for comparison. The obtained results showed that the SVM with Gaussian-kernel provided better discrimination accuracy than that obtained by the conventional discriminant algorithm used the discrete NDVI-based aggregation data.

Another work, [52] employed the RF for weeds(Zea mays, Convolvulus arvensis, Rumex, Cirsium arvense) recognition in a maize crop using near-infrared snapshot mosaic hyperspectral imagery. Experiments were performed using three distinct combinations of features and compared with the KNN. Results presented a significant overall better performance of the RF model.

VI. Crop and Fruit Detection

Crop and fruits detection is a kind of crop prediction, but it is based on the detection of the presence of fruits from images. It aims at providing information to farmers to optimise the economic benefits and make plans for their agricultural work, in addition, it helps to adjust management practices before harvesting and to decide proper investments in advance because it offers an early estimation of yields and fruit growth.

The study presented in [53] investigated the suitability of applying a deep learning algorithm and CNN for strawberry (mature and immature) recognition based on greenhouse images. The study tried to propose solution for training and learning a CNN using a small set of data, it uses 373 images for training and for testing. The developed CNN achieved a good precision.

[54] Proposed an automatic, efficient and low-cost fruit count method of coffee branches using computer vision. Based on the state of coffee fruits (harvestable, not harvestable, fruits with disregarded maturation stage), the method aimed at calculating the number of these fruits, and estimated the weight and the maturation percentage of the coffee fruits. After the process of image pre-processing, three classifiers were implemented to perform tasks of the detection, classification and fruits' count; Bayes classifier; KNN; and SVM classifier. After that, and according to the experimental results, the authors have selected the SVM to validate their model because it outperformed the Bayes and KNN.

[55] Suggested a machine vision system for detecting cherry tree branches in planar architecture using BC, where the objective of the work was to reduce labour requirements in manual harvesting and handling operations. The system segments and detects cherry branches only if the intermittent segments of branches were visible. The BC classifies image pixels into four classes: (branch, background, cherry, leaf), and it achieved good accuracy in identifying branch pixels.

[56] Presented a detection method of tomatoes based on Expectation Maximisation (EM) and remotely sensed RGB images which were captured by an UAV. It employed several techniques of data mining, it first started by clustering technique using Bayesian information criterion to define the optimal number of clusters for the image. Then, it used K-means to carry out the spectral clustering, where the EM and Self Organising Map (SOM) algorithms are utilised to classify pixels into two groups (tomatoes, non-tomatoes). The results indicated that EM outperformed SOM and K-means.

[57] Developed an early yield mapping system for the detection of immature green citrus based on machine vision. Like all other relative studies, and after images pre-processing and features extractions, SVM was applied to retrieve the immature citrus. The accuracy was achieved a good value.

Another model, called DeepFruits, was proposed in [58] for sweet pepper detection from imagery data, RGB colour and Near-Infrared. This work employed a CNN and adopted for the Faster Region-based CNN (Faster R-CNN). The model was trained, and its performance through the measure of precision and recall was promising.

[59] Presented computer vision algorithms to detect and count immature peach from colour images acquired in natural illumination conditions using different parametric and non-parametric classifiers: (ANN, discriminant analysis, Naive Bayes, KNN, trees classifier, regression trees, SVM).According to the obtained overall performances of classifiers, it has concluded that the parametricity was not a significant factor in detection performance with respect to classifier type. The worst detection rate was obtained by the Naive Bayes classifier, this because it is known by its poor capability to deal with complex interactions between features.

VII. Clustering Techniques for Crop Yield

Clustering techniques are not widely employed in DA, few efforts have investigated the potential of such techniques. Although, there are two sub-classes of clustering for crop yields: modelling and delineation of management zones.

A. Modelling

[60] Proposed a methodology for data mining and knowledge discovery in smart farming called "KDLC Knowledge Discovery Life Cycle". It consisted of 6 activities that accompany and help the user throughout the KDD process. It combined supervised inductive rule learning AQ5c and unsupervised BC to construct a multi-strategy knowledge discovery approach. This approach started by analysing datasets utilising BC to discover the important taxonomic classes, these later were represented as new attributes in an expanded representation space via constructive induction mechanism, which was then exploited to learn concepts, relationships, and rules that characterise knowledge in the data space.

[61], [62] Introduced a modelling approach of Fuzzy Cognitive Map (FCoM) to help on the decision-making. It utilised the soft computing technique of FCoM to handle initial knowledge. The FCoM was enriched with an application of unsupervised learning algorithm the nonlinear Hebbian learning (NHL), to characterise the data into two production yield categories, and to provide a description of cotton yield trend and behaviour. This methodology was related to set of

experts who supervise the system and expect its behaviour, it extracted the knowledge from those experts and exploited their experiences. Experts had have knowledge on which elements of the system influence other elements; they could define the positive & negative effects of one concept on the others, with a fuzzy degree of causation for the corresponding concepts. The NHL algorithm is proposed to train FCoM; it is introduced to overcome inadequate knowledge of experts and/or non-acceptable FCoM simulation results and to adapt weights. The accuracy of the FCoM model and the implementation of the NHL algorithm for (low, high) yield categories for the three respective years of 2001, 2003 and 2006 has achieved an acceptable success rates.

[63] Presented a hierarchical grading method applied to JonaGold apples. After the acquirement of images that cover the whole surface of the fruits, images then were segmented to extract the defected features. Unsupervised learning K-means clustering algorithm was used with number of clusters fixed by preliminary studies to 16 for the variety of apples. To grade the fruits, a principal component analysis was used with the 16 first principal components which 97% of the whole variations. The global correct clustering was achieved an acceptable rate.

To classify plants (sunflower, red-root pig-weed, soybean, velvet-leaf plants), soil(bare clay) and residue regions of interest (ROI) (corn and wheat residues), [64] proposed to use an intensified fuzzy clusters FCM and Gustafson-Kessel (GK) in conjunction with Fuzzy excess red (ExR) and excess green (ExG) indices, all enhanced with Zadeh's (Z) fuzzy intensification technique, to classify hidden and prominent (ROI) in colour images. The paper concluded that the Zadeh GK algorithm was very useful for crop management, mapping, remote sensing, pests and weeds control.

B. Delineation of management zones

Usually, farmers split their agricultural land into fields for several reasons; to variate their crops and make crop-rotation practices, to facilitate the management tasks and to create their yields maps to help them in enhancing their crop yields. This process is called delineation of management zones (DMZ). DMZ of yields is an important task for crop monitoring since it aims at determining zones of low-or-high yields, and to find out reasons behind low yield fields, hence, to propose solutions to increase yields of the fields known by their low productivity.

From the DA view, given an area divided into zones, DMZs process tries to find adjacent zones which exhibit similar characteristics or homogeneous with other zones, at the same time are heterogeneous and have different characteristics with other zones. This can be translated using data mining vocabulary by clustering. Based on the literature review, and for both techniques for delineation of yields' zone management, the most used technique is k-means clustering (non-hierarchical method).

Figure 4 depicts the general process of delineation of management zones followed by the majority of proposed works.

[65] Presented a study that aimed to emphasise the possibility of using the hierarchical method as complementary to

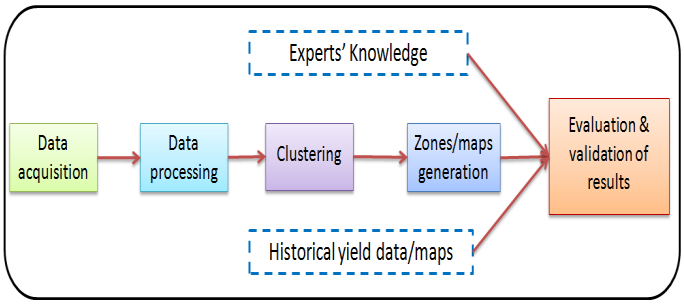

Fig. 4. The delineation management zones process.

the non-hierarchical clustering method in order to estimate the number of management zones of a given field which can be used as an input for the non-hierarchical method. For this end, it collected data from yield monitoring during three seasons for three different crops (Spring barley; oil-seed rape; winter wheat). After pre-processing data, and in order to approximate the management zones within the given field, 31 monitoring points were selected across the field and data from all data-sets were allocated to these points to be used as an input data-set for the analysis. Then, the cubic clustering criterion (CCC) has been used to approximate the statistically significant number of clusters for both Ward's method (hierarchic method) and k-means method. The conducted clustering uses in the first step the hierarchic method (Ward's method) in order to explore the data based on the homogeneity which allows to identify the homogeneous places, followed by the non-hierarchic method (k-means clustering) which creates clusters which are as heterogeneous as possible. In other words, the hierarchical clustering procedure used to identify the number of clusters and the non-hierarchical procedure used to identify the members of each cluster and its characteristics. The results of this study demonstrated that the use of both hierarchical and non-hierarchical clustering methods was beneficial for the determination of the management zones from yield maps.

Using method similar to the previous one, [66] evaluated the capability of identifying the yield potential zones based on historical yield maps, and tested the approach over the growing extent of input data. It used data from six growing seasons for (spring barley, winter wheat, spring barley, spring oil-seed rape, winter wheat, spring barley) and selected 67 monitoring points. Results demonstrated that despite the complexity of data from commercial combine monitoring systems, its utilisation enables determining the zones.

K-Means algorithm groups data by the similarity of their values, using some distance measures such as the Euclidean distance, but for the task of delineation, data are also characterised by geographical coordinates associated with each sample, which leads to the idea of grouping data using other factors like the geographical distance determined between samples. However, [67] argued that data are also associated with a location, which implies an underlying degree of vagueness caused by some factors like the accuracy of the GPS capture device. Furthermore, It makes the grouping not reliable

if obtained as hard clusters, like the k-means algorithm did. Therefore, for DMZ applications, where samples may belongs to more than one cluster with certain degree of membership, fuzzy clustering algorithm and clustering methods that allow such imprecision are very useful.

[68] Proposed an approach for the delineation of PMZ for crop management (Zea mays L.) that expresses the productive potential of the soil within a field, using farmer's expert knowledge and data on yields. The used data contained (yield maps, soil electrical conductivity, remote sensing multispectral indices, topographical data). It implemented FCM to create different alternatives of PMZ, then it has taken into consideration the farmers' expert knowledge to improve the resulting PMZs that best fitted to the yield spatial variability pattern.

In the same way, numerous works used either KMeans or FCMs for DMZs of yields; among others, [69] proposed to delineate and manage grapevines zones using fuzzy clustering techniques, in addition to yield and grape composition data in addition to soil properties. [70] Applied FCMs algorithm in geo-referenced yield and grain moisture data-set to obtain the optimal number of homogeneous zones. [71] Employed different analysis techniques: k-means, a multivariate variance and discriminant analysis in an attempt to examine a framework for cotton's zones delineation. [72] Employed FCM algorithm to develop a software called ZoneMAP for an automated delineation of management zones using field data provided by users and satellite imagery. [73] Used FCM to delineate management zones considering historical yield data from corn-soybean rotation crops, identifying the spatial association of the obtained maps with soil maps. [74] Used FCM clustering to delineate MZs on NDVI, salinity and yield data in cotton.

VIII. Discussion

In this section, we will draw some remarks regarding the application of data (mining/analytics) in DA and their extent of use of big data concepts.

From the data mining perspectives, methods analysed in this review, regardless the examined discipline as described in Figure 2, employ the same process: data/image acquisition, then processing and features selection, followed by the clustering or classification task; which is considered as a typical data mining application. Of course, each discipline has its specificities, and each problem within the same discipline has its characteristics, obstacles, and has a special kind of data with variation in type and volume, thence each problem requires a particular solution and adequate algorithm considering the above-mentioned differences.

Therefore, it is not wise to recommend a solution or an algorithm for a problem, to prefer a solution than another or to judge an algorithm that it is better than another one. Although the application of machine learning algorithms in DA is intuitive, challenges encounter this application and the performance of these algorithms are non-trivial. It is well-known that machine learning algorithms are sensitive to data used to train them. As we have seen, AI and computer vision have also emerged in DA in conjunction with machine learning.

Images, sensors and satellites data types require a pre-processing and segmentation before using them which has been performed using AI techniques. The quality of the result of image processing depends on the quality of images, which in its turn, depends on the acquisition step, the used devices to capture them, to the lightning background and other factors. It is obvious that these facts influence the performance of the machine learning algorithm. This reality is highly manifested in the case of delineation of management zones, where the process is almost the same, even for the choice of the clustering algorithm; the differences were in the choice of data type, choices for algorithms to process data, the number of features, etc. The same remarks are relevant to the other applications like crops classification, crops protection and detection.

In the case of yield prediction; we have found two kinds of forecasting depend on the nature of the used data:

- A yield forecast system based on historical data provides a pre-season estimation of the yield, and even before the beginning of the crop season which can give time and capability to farmers to make decision on which strategy to follow and which step to change to enhance the crop yield ahead in the crop cycle, like choosing seeds and crop type that match more the climate and weather variations, soil state, etc. In addition, it can work in large fields without any additional cost. The performance of machine learning algorithms for this type of forecasting depends on the (quality & quantity) of data which will feed the algorithms.
- On the other hand, forecasting system employed the other types of data regarding images issued by satellite, cameras, scanner, sensors, allows for on-season estimation, at the beginning of the growing season, at the middle or at the end of season just before the harvest, since it requires treating images for the crops upon which it performs the estimation. So that, it can provide too, near real-time insights into the state of crops and other problems like diseases. This system can work on small to large fields but with additional cost for images' acquisition. For example, it is known that satellite imagery is expensive and not within the reach of all farmers, in addition, most of the reviewed works showed that the use of the NDVI data is time-consuming to be acquired and processed. The performance of machine learning algorithms using these data depends on the quality of images, to the image processing and features selection process.

Another observation from the analysed works revealed that the ANN with its numerous implementations had dominated the prediction tasks, the SVM for the classification and detection, and the K-means for the clustering. It is worth to notice the emergent use of deep learning and CNN for the most recent works on detection, classification and prediction, which in fact presents impressive results. This is because of the

availability of sophisticated computational hardware like GPU which allows for the processing of voluminous and complex data.

From the big data perspectives, an application is said to be applied big data concept if the used data-set can be described by four primary characteristics: volume, velocity, variety and veracity, commonly know by 4Vs.

- Volume (V1): The size of data collected for analysis;
- Velocity (V2): The frequency of collecting data. Data are accumulated in real-time and / or at a rapid pace, occasionally, yearly...;
- Variety (V3): The nature of data used and its sources: historical or image, or a combination of both. For example it can have a multi-sources from reports, videos,images, remote and sensing data...;
- Veracity (V4): The quality, reliability and the accuracy of the data;

It is obvious that more data is complex regarding its 4Vs more their analysis is complex too, so, considering the employment of the four metrics of big-data in DA we draw the following remarks:

- Velocity: it is remarked that many works do not mentioned the frequency of accumulating their data. Generally speaking, in DA the frequency of collecting data depends on the nature of data itself and to the problem for which data were collected. Some applications need a real-time data and others do not. Data-set for the prediction of crop yields used historical data have low velocity comparing to data-set collected for the protection and disease detection of crops using sensors or other type of image data which require a day per day control and hence, real-time data.
- Variety: most of the examined works used multi-source of data and in many cases a combination of historical and image data were exploited.
- Veracity: it is observed that most of the used data-sets needed to be cleaned and pre-processed, and that more the variety and velocity is high in the used data-set more its veracity is high too. [75] Stated that: *increased variety and high velocity hinder the ability to cleanse data before analysing it and making decisions, magnifying the issue of data 'trust'*.

Considering the volume metric of big data, it is noticed that the volume of data-sets used by most of the analysed works does not meet the standard of big data. This reality is due to the following points which are considered are barriers against the full utilisation of big data in DA:

- DA is a new concept for farmers, and it is not applied until very recently. Before, farmers are not interested and motivated to collect data except for few cases like yields and weather conditions, or for statistics' purposes;
- Data-sets are usually collected from individual small farms and laboratories which cannot allow to generate a big mass of data.
- Big mass of data can be generated by big farms which usually belong to big companies, so that for security reasons and for competitions too, these companies avoid and do not prefer to share their data or to publish it.

Besides, the volume metric is highly depends on the nature of data, application employed satellite data for example is expected to use a high volume of data due to the size of pictures. It can depends too to the nature of the problem, size of data for yield prediction problems has tendency to be smaller than those used for crop protection. Moreover, with today sophisticated machines and algorithms, the volume of data is not the most important factor to worry about and it is not the most critical challenge. Veracity, velocity and variety are more essential and crucial because they add more complexity to the analysis and to the pre-processing of the data.

For these reasons, we are not considerate the volume criterion, which indicates the size of data-set and shows how much is it big. In addition, the bigness is not entirely about the size of data set, but also about the other three elements.

Table I, resumes a set of representative papers in DA according to their usage of big data. For each paper, we identify the type, the size, the heterogeneity of data used, and the frequency of its collection. In addition, we consider the number and type of machine learning algorithms used, the complexity of the proposed analyse algorithms and the used device to collect data.

In Table I:

No: data were clear and all samples have been used;
Yes: data were cleaned and filtered and some samples were not considered because of abnormalities, inconsistencies or duplication and for other reasons;
n: number of training simple or data points;
n_{sv}: number of support vectors;
P: number of features;
n_{trees}: number of trees;
c : number of cluster;
d : number of dimension;

i : number of iterations;
L: number of hidden layers;
TQ is the size of input feature map; spatial, two/three-dimensional kernels are of size (tq);
nl_i :number of neurons at layer i;
ep: number of epochs.

From Table I, we can extract three classes of applications according to their usage intensity of the big data metrics: Full usage, light usage, non usage.

- Full usage: are applications that fully employ big data by all of its elements;
- Light usage: are applications that partially employ big data elements;
- Non usage: are applications that do not have any kind of use of big data concept and elements.

IX. Conclusion

Digital agriculture is in the way to re-shape the farming practices by making it more controllable and accurate; its key component is the use of ICT, sensors, GPS and other technologies for the benefit of farmers and the enhancement of its crops.

Machine learning and data mining techniques are reliable techniques for analysing data and exploring new information from these data. On the other hand, big data adds additional support to DA by discovering further insights from the collected data in order to solve farming problems and inform farming decisions.

This survey presented a systematic review of the application of machine learning, data mining and analytics in the agricultural sector. It was first exhibited the process of crops management and its different parts, where we are focused on the crop yield monitoring. Then, for this later, it has provided a classification of the several employment of data mining techniques into this field. For each class of the classification, a set of existing works have been reviewed to demonstrate the machine learning method applied and for which purpose.

After that, the survey discussed the applicability of big data concepts, and it demonstrated that DA is on the road to exploit the full potential of big data concepts. This will open the gate to new opportunities of investment into these fields and will allow for a very different management way of crops, it promised for advanced scientific discoveries and innovative solutions to more complicated agricultural issues. In addition, it will provide farmers with new insights into how they can grow crops more efficiently.

The survey established that despite all the advantages gained from DA, there are several challenges and obstacles need to be surmounted in order to make from DA a real data-driven solution, among them lack of data because of several reasons like data ownership rights, data (or agricultural knowledge) providers needs guarantees for both their investments (money) in DA and for their security (competitions and many other facts), in addition, they usually need to acquire new skills to understand new technologies, which means an additional investment in term of time and effort.

To conclude, we say that "if energies are the soul of machines, then data are the spirit of algorithms".

References

[1] K. Poppe, S. Wolfert, C. Verdouw, and T. Verwaart, "Information and communication technology as a driver for change in agri-food chains," EuroChoices 12, 60–65, Tech. Rep., 2013.

[2] K. Soma, M. Bogaardt, K. Poppe, S. Wolfert, G. Beers, D. Urdu, M. P. Kirova, C. Thurston, and C. M. Belles, "Research for agri committee - impacts of the digital economy on the food chain and the cap. policy department for structural and cohesion policies," European Parliament. Brussels, Tech. Rep., 2019.

[3] "Preparing for future akis in europe," European Commission. Brussels, Tech. Rep., 2019.

[4] S. Wolfert, C. Verdouw, and M. Bogaardt, "Big data in smart farming – a review," *Agricultural Systems*, vol. 153, pp. 69–80, 2017.

[5] T. Stombaugh and S. Shearer, "Equipment technologies for precision agriculture," *Journal of Soil and Water Conservation*, vol. 55, no. 1, pp. 6–11, 2000.

[6] G.Pelletier and S. Upadhyaya, "Development of a tomato load/yield monitor," *Computers and Electronics in Agriculture*, vol. 23, pp. 103–117, 1999.

[7] D. Elavarasan, D. Vincent, V. Sharma, A. Zomaya, and K. Srinivasan, "Forecasting yield by integrating agrarian factors and machine learning models: A survey," *Computers and Electronics in Agriculture*, vol. 155, pp. 257–282, 2018.

[8] D. Patricio and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," *Computers and Electronics in Agriculture*, vol. 153, pp. 69–81, 2018.

[9] A. Kamilaris, A. Kartakoullis, and F. Prenafeta-Boldu, "A review on the practice of big data analysis in agriculture," *Computers and Electronics in Agriculture*, vol. 143, pp. 23–37, 2017.

[10] J. Behmann, A.K.Mahlein, T.Rumpf, C.Romer, and L. Plumer, "A review of advanced machine learning methods for the detection of biotic stress in precision crop protection," *Journal of Precision Agriculture*, vol. 16, pp. 239–260, 2014.

[11] A. Mucherino, P. Papajorgji, and P. M. Pardalos, "A survey of data mining techniques applied to agriculture," *Journal of Operational Research*, vol. 9, no. 2, pp. 121–140, 2009.

[12] S. Sabzi and Y. Abbaspour-Gilandeh, "Using video processing to classify potato plant and three types of weed using hybrid of artificial neural network and particle swarm algorithm," *Measurement*, vol. 126, pp. 22–36, 2018.

[13] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778–782, 2017.

[14] M. Dyrmann, H. Karstoft, and H. Midtiby, "Plant species classification using deep convolutional neural network," *Biosystems engineering*, vol. 151, pp. 72–80, 2016.

[15] S. Contiu and A. Groza, "Improving remote sensing crop classification by argumentation-based conflict resolution in ensemble learning," *Expert Systems With Application*, vol. 64, pp. 269–286, 2016.

[16] A. Formaggio, M. Vieira, and C. Renno, "Object based image analysis (obia) and data mining (dm) in landsat time series for mapping soybean in intensive agricultural regions," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium.*, Munich Germany, Jul. 2012, pp. 2257–2260.

[17] J. Arribas, G. Sanches-Ferrero, G. Ruiz-Ruiz, and J. Gomez-Gil, "Leaf classification in sunflower crops by computer vision and neural networks," *Computers and Electronics in Agriculture*, vol. 78, no. 1, pp. 9–18, 2011.

[18] F. Ahmed, H. Al-Mamun, H. Bari, E. Hossain, and P. Kwan, "Classification of crops and weeds from digital images: a support vector machine approach," *Crop Protection*, vol. 40, pp. 98–104, 2012.

[19] J. Ambuel, T. S. Colvin, and D. L. Karlen, "A fuzzy logic yield simulator for prescription farming," *Transactions of the ASAE*, vol. 37, no. 6, pp. 1999–2009, 1994.

[20] S. Drummond, K. A. Sudduth, and S. J. Birrell, "Analysis and correlation methods for spatial data," in *ASAE Paper No. 95–1335. St. Joseph, Mich.: ASAE.*, 1995, pp. 350–359.

[21] S. Drummond, K. Sudduth, A. Joshi, S. Birrell, and N. Kitchen, "Statistical and neural methods for site-specific yield prediction," *Transactions of the ASAE*, vol. 46, no. 1, pp. 5–14, 2003.

[22] B. Ji, Y. Sun, S.Yang, and J. Wan, "Artificial neural networks for rice yield prediction in mountainous regions," *Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science.*, vol. 145, no. 3, pp. 249–261, 2007.

[23] G. RuB, M. S. R. Krus and, and P. Wagner, "Optimizing wheat yield prediction using different topologies of neural networks," in *Proceedings of IPMU-08*, 2008, pp. 576–582.

[24] G. RuB, "Data mining of agricultural yield data: A comparison of regression models," in *Industrial Conference on Data Mining ICDM 2009: Advances in Data Mining. Applications and Theoretical Aspects, P. Perner (Ed.), Lecture Notes in Artificial Intelligence 6171, Berlin, Heidelberg, Springer*, 2009, pp. 24–37.

[25] S. Fukuda, W. Spreer, E. Yasunaga, K. Yuge, V. Sardsud, and J. Muller, "Random forests modelling for the estimation of mango (mangifera indica l. cv.chok anan) fruit yields under different irrigation regimes," *Journal of Agricultural Water Management*, vol. 116, no. 3, pp. 142–150, 2013.

[26] I. Oliveira, R. Cunha, B. Silva, and M. Netto, "A scalable machine learning system for pre-season agriculture yield forecast," in *the 14th IEEE eScience Conference*, 2018.

TABLE I
DA APPLICATIONS AND THEIR USAGE OF BIG DATA CONCEPTS.

Ref	Volume	Velocity	Veracity	Variety	ML	Complexity	Device	Task
[18]	224 images	/	No	Image Data digital images	SVM	$O(n^2p + n^3) + O(n_{sv}p)$	Digital camera	Classification
[73]	3*2 years of data monitoring	1 year	/	Sensor data: soil properties	Fuzzy C-means	time: $O(ndc^2i)$ space:$O(nd + nc)$	Pressure-based: AgLeader Ames,IA	Clustering
[15]	/	/	No	Satellite data: Images in GeoTiff	EL (DT+ SVM+ ANN)	$O(n^2p) + O(p) + O(n^2p + n^3)$ $+O(n_{sv}p)+$ $O(epn(nl_1nl_2 + nl_2nl_3 + ...)+$ $O(pnl_1 + nl_1nl_2 + nl_2nl_3 + ...)$	Satellite	Classification
[20]	3000 for topographical data, 3120 data points for the other types	1 year for crop yield and soil's composite 1 day for climat	No	Image and sensor data: Soil properties Topographic crop yield climatological	ANN	$O(epn(nl_1nl_2 + nl_2nl_3 + ...)+$ $O(pnl_1 + nl_1nl_2 + nl_2nl_3 + ...)$	Camera Nikon Topgun A200LG electromagnetic induction sensor Yield sensor and GPS	Prediction
[30]	/	/	Yes	All types of data: yield, soil information Geo-physical Remote sensed Climate	RF	$O(n^2pn_{trees}) + O(pn_{trees})$	Yield monitor soil-maps, EM gamma survey MODIS NDVI	Prediction
[66]	229	1 year	Yes	Historical data: Crop yield	K-means	$O(ncdi)$	/	Clustering
[14]	10413	/	/	Image data: Digital images	CNN	$O(TQtq)$	Cell phone	Classifiation
[29]	/	1 year	No	Historical data: Crop yield soil parameters	ELM	$O(L^3 + L^2n)$	/	Prediction
[59]	96	1year	Yes	Image data: Digital images	SVM,ANN,NB KNN DT Discriminant analysis	Discriminant analysis: $O(np^2)$ NB: $O(np) + O(p)$	Camera Nikon CoolpixL22	Classification
[36]	8945	multi-spectral image: 8 days interval for 30 times a year	Yes	Satellite and sensor data: surface reflectance land surface temperature land cover	Gaussian CNN	$O(TQt2)$	MODIS satellite	Prediction

[27] A. Gonzalez-Sanchez, J. Frausto-Solis, and W. Ojeda-Bustamante, "Predictive ability of machine learning methods for massive crop yield prediction," *Spanish Journal of Agricultural Research*, vol. 12, no. 2, pp. 313–328, 2014.

[28] J. Jeong, J. Resop, N. Mueller, D. Fleisher, K.Yun, E. Butler, D. Timlin, K. Shim, J. Gerber, V. Reddy, and S. Kim, "Random forests for global and regional crop yield predictions," *PLoS ONE*, vol. 11, no. 6, 2016.

[29] L. Kouadio, R. Deo, V. Byrareddy, J. Adamowski, S. Mushtaq, and V. P. Nguyen, "Artificial intelligence approach for the prediction of robusta coffee yield using soil fertility properties," *Computers and Electronics in Agriculture*, vol. 155, pp. 324–338, 2018.

[30] P. Filippi, E. Jones, T. Bishop, N. Acharige, S. Dewage, L. Johnson, S. Ugbaje, T. Jephcott, S. Paterson, and B. Whelan, "A big data approach to predicting crop yield," in *Proceedings of the 7th Asian-Australasian Conference on Precision Agriculture 16–18 October 2017, Hamilton, New Zealand.*, 2017.

[31] X. Ye, K. Sakai, L. Garciano, S. Asada, and A. Sasao, "Estimation of citrus yield from airborne hyperspectral images using a neural network model," *Ecological Modelling*, vol. 198, no. 3-4, pp. 426–432, 2006.

[32] X. Pantazi, D. Moshou, T. Alexandridis, R. Whetton, and A. Mouazen, "Wheat yield prediction using machine learning and advanced sensing techniques," *Journal of Computers and Electronics in Agriculture*, vol. 121, pp. 57–65, 2016.

[33] R.Ji, J.Min, Y.Wang, H.Cheng, H.Zhang, and W.Shi, "In-season yield prediction of cabbage with a hand-held active canopy sensor," *Sensors*, vol. 17, no. 10, 2017.

[34] H. Cheng, L. Damerow, Y. Sun, and M. Blanke, "Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks," *Journal of imaging*, vol. 3, no. 1, 2017.

[35] A. Wang, C. Tran, N. Desai, D. Lobell, and S. Ermon, "Deep transfer learning for crop yield prediction with remote sensing data," in *Proceedings of the COMPASS'18, Proceedings of the 1st ACM SIGCAS conference on Computing and Sustainable Societies. Menlo Park and San Jose, CA, USA. June 20-22*, 2018.

[36] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep gaussian process for crop yield prediction based on remote sensing data," in *the Thirty-First AAAI Conference on Artificial Intelligence. AAAI Publications*, 2017, pp. 4559–4566.

[37] K. Kuwata and R. Shibasaki, "Estimating crop yields with deep learning and remotely sensed data," in *Geoscience and Remote Sensing Symposium, IEEE International*, 2015, pp. 858–861.

[38] S. Mohanty, D. Hughes, and M. Salathe, "Using deep learning for image-based plant disease detection," *Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science.*, vol. 7, pp. 1–10, 2016.

[39] D. Al-Bashish, M. Braik, and S.Bani-Ahmed, "Detection and classification of leaf disease using k-means-based segmentation and neural networks-based classification," *Information Technology. Asian Network of scientific information*, vol. 10, no. 2, pp. 267–275, 2011.

[40] Q. Yao, Z. Guan, Y. Zhou, J. Tang, Y. Hu, and B. Yang, "Application of support vector machine for detecting rice diseases using shape and color texture features," in *International Conference on Engineering Computation, IEEE computer society. 2-3 May 2009 Hong Kong, China*, 2009, pp. 79–83.

[41] K. Huang, "Application of artificial neural network for detecting phalaenopsis seedling diseases using color and texture features," *Computers and Electronics in Agriculture*, vol. 57, no. 1, pp. 3–11, 2007.

[42] Y. Tian, T. Li, C. Li, Z. Piao, G. Sun, and B. Wang, "Method for recognition of grape disease based on support vector machine," *Transaction. CSAE*, vol. 23, no. 6, pp. 175–180, 2007.

[43] S. Sladojevic, M. Arsenovic, A. A. D. Culibrk, and D. Stefanovic, "Deep neural networks based recognition of plant diseases by leaf image classification," *Computational Intelligence and Neuroscience*, vol. 2016, 2016.

[44] Z. Liu, H. Wu, and J. Huang, "Application of neural networks to discriminate fungal infection levels in rice panicles using hyperspectral reflectance and principal components analysis," *Computers and Electronics in Agriculture*, vol. 72, no. 2, pp. 99–106, 2010.

[45] M. El-Telbany and M. Warda, "An empirical comparison of tree-based learning algorithms: An egyptian rice diseases classification case study," *International Journal of Advanced Research in Artificial Intelligence*, vol. 5, no. 1, 2016.

[46] K. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Computers and Electronics in Agriculture*, vol. 145, pp. 311–318, 2018.

[47] B. Liu, Y. Zhang, D. He, and Y. Li, "Identification of apple leaf diseases based on deep convolutional neural networks," *Symmetry*, vol. 10, no. 1, 2018.

[48] K. Yamamoto, T. Togami, and N. Yamaguch., "Super-resolution of plant disease images for the acceleration of image-based phenotyping and vigor diagnosis in agriculture," *Sensors*, vol. 17, no. 11, 2017.

[49] E. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture. In press*, 2018.

[50] A. Cruz, A. Luvisi, L. D. Bellis, and Y. Ampatzidis, "X-fido: An effective application for detecting olive quick decline syndrome with deep learning and data fusion," *Frontiers Plant Science*, vol. 8, 2017.

[51] S. Akbarzadeh, A. Paap, S. Ahderom, B. Apopei, and K. Alameh, "Plant discrimination by support vector machine classifier based on spectral reflectance," *Computers and Electronics in Agriculture*, vol. 148, pp. 250–258, 2018.

[52] J. Gao, D. Nuyttens, P. Lootens, Y. He, and J. Pieters, "Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery," *Biosystems Engineering*, vol. 170, pp. 30–50, 2018.

[53] H. Habaragamuwa, Y. Ogawa, T. Suzuki, T. Masanori, and O. Kondo, "Detecting greenhouse strawberries (mature and immature), using deep convolutional neural network," *Engineering in Agriculture, Environment and Food*, vol. 11, no. 3, pp. 127–138, 2018.

[54] P. Ramos, F. Prieto, E. Montoya, and C. Oliveros, "Automatic fruit count on coffee branches using computer vision," *Computers and Electronics in Agriculture*, vol. 137, pp. 9–22, 2017.

[55] S. Amatya, M. Karkee, A. Gongal, Q. Zhang, and M. Whiting, "Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting," *Biosystems Engeneering*, vol. 146, pp. 3–15, 2015.

[56] J. Senthilnath, A. Dokania, M. Kandukuri, K. Ramesh, G. Anand, and S. Omkar, "Detection of tomatoes using spectral-spatial methods in remotely sensed rgb images captured by uav," *Biosystems Engineering*, vol. 146, pp. 16–32, 2016.

[57] S. Sengupta and W. Lee, "Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions," *Biosystems Engineering*, vol. 117, pp. 51–61, 2014.

[58] I. Sa, Z. Ge, F. D. B. Upcroft, T. Perez, and C. Mccool, "Deepfruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, no. 8, 2016.

[59] F. Kurtulmus, W. Lee, and A.Vardar, "Immature peach detection in colour images acquired in natural illumination conditions using statistical classifiers and neural network," *Precision Agriculture*, vol. 15, no. 1, pp. 57–79, 2014.

[60] S. Lee and L. Kerschberg, "Methodology and life cycle model for data mining and knowledge discovery in precision agriculture," in *the IEEE International Conference on Systems, Man and Cybernetics, Vol. 3*, 1998, pp. 2882–2887.

[61] E. Papageorgiou, A. Markinos, and T. Gemptos, "Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application," *Applied Soft Computing*, vol. 11, no. 4, pp. 3643–3657, 2011.

[62] ——, "Application of fuzzy cognitive maps for cotton yield management in precision farming," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12 399–12 413, 2009.

[63] V. Leemans and M. Destain, "A real-time grading method of apples based on features extracted from defects," *Journal of Food Engineering*, vol. 61, no. 1, pp. 83–89, 2004.

[64] G. Meyer, J. Neto, D. Jones, and T. Hindman, "Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images," *Computers and Electronics in Agriculture*, vol. 42, no. 3, pp. 161–180, 2004.

[65] J. Galambosova, V. Rataj, R. Prokeinova, and J. Presinska, "Determining the management zones with hierarchic and non-hierarchic clustering methods," *Special Issue in Research in Agriculture Engineering*, vol. 60, pp. 44–51, 2014.

[66] M. Ingeli, J. Galambosova, R. Prokeinova, and V. Rataj, "Application of clustering method to determine production zones of field," *Acta Technologica Agriculturae*, vol. 18, no. 2, pp. 42–45, 2015.

[67] E. Speranza, R. Ciferri, C. Grego, and L. Vicente, "A cluster-based approach to support the delination of management zones in precision agriculture," in *IEEE 10 th International Conference on eScience*, 2014.

[68] J. Martinez-Casasnovas, A. Escola, and J. Arno, "Use of farmer knowledge in the delineation of potential management zones in precision agriculture: A case study in maize (zea mays l.)," *Agriculture*, vol. 84, no. 8, 2018.

[69] A. Tagarakis, V. Liakos, S. Fountas, S. Koundouras, and T. Gemtos, "Management zones delineation using fuzzy clustering techniques in grapevines," *Precision Agriculture*, vol. 14, no. 1, pp. 18–39, 2013.

[70] L. Vendrusculo and A. Kaleita, "Modeling zone management in precision agriculture through fuzzy c-means technique at spatial database." in *Proceedings of the 2011 ASABE Annual International Meeting Sponsored by ASABE. Gault House, Louisville, Kentucky. August 7-10*, 2016, pp. 350–359.

[71] J. Ping, C. Green, K. Bronson, R. Zartman, and A. Dobermann, "Delineating potential management zones for cotton based on yields and soil properties," *Soil Science*, vol. 170, no. 5, pp. 371–385, 2005.

[72] X. Zhang, L. Shi, X. Jia, G. Seielstad, and C. Helgason, "Zone mapping application for precision farming: a decision support tool for variable rate application," *Precision Agriculture*, vol. 11, no. 2, pp. 103–114, 2010.

[73] A. Brock, S. Brouder, G. Blumhoff, and B. Hofmann, "Defining yield-based management zones for corn-soybean rotations," *Agronomy Journal*, vol. 97, no. 4, pp. 1115–1128, 2005.

[74] L. Yan, S. Zhou, W. Cifang, L. Hongyi, and L. Feng, "Classification of management zones for precision farming in saline soil based on multi-data sources to characterize spatial variability of soil properties," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 23, no. 8, pp. 84–89, 2007.

[75] B. Feldman, E. Martin, and T. Skotnes, "Big data in healthcare hype and hope, october 2012.dr. bonnie 360, 2012." 2012.