 doi:10.20944/preprints202202.0345.v1

Review

Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review

Ania Cravero [1,*], Sebastian Pardo [1], Samuel Sepúlveda [1] and Lilia Muñoz [2]

[1] Department of Computer Science and Informatics, Center for Software Engineering Studies, Universidad de La Frontera, Temuco, Chile; s.pardo02@ufromail.cl; samuel.sepulveda@ufrontera.cl

[2] Faculty of Computer Systems Engineering es, Universidad Tecnológica de Panamá, Panama; lilia.munoz@utp.ac.pa

* Correspondence: ania.cravero@ufrontera.cl

Abstract: Agricultural Big Data is a set of technologies that allows responding to the challenges of the new data era. In conjunction with machine learning, farmers can use data to address different problems such as farmers' decision-making, crops, weeds, animal research, land, food availability and security, weather, and climate change. The purpose of this paper is to synthesize the evidence regarding the challenges involved in implementing machine learning in Agricultural Big Data. We conducted a Systematic Literature Review applying the PRISMA protocol. This review includes 30 papers, published from 2015 to 2020. We develop a framework that summarizes the main challenges encountered, the use of machine learning techniques, as well as the main technologies used. A major challenge is the design of Agricultural Big Data architectures, due to the need to modify the set of technologies adapting the machine learning techniques, as the volume of data increases.

Keywords: Big Data; Machine Learning; Agriculture; Challenges; Systematic Literature Review.

1. Introduction

To meet global food demand by 2050 requires an increase in food production from 25% to 70% [1]. Because of this increase, food production per hectare needs to double by the time the world population stabilizes around 2100 (United Nations 2019). Food security is a fundamental global need, which is threatened by several factors such as population growth, shrinking arable land, climate change, food waste, and consumer preference for animal protein [2]. Increasing agriculture or food production rapidly to meet the growing demand for food supply is not an easy task. Several factors contribute to this problem, such as current agricultural practices, poor storage, markets, and changing scenarios [3].

For offering sustainable agricultural production, it is necessary to use cutting-edge technologies such as Blockchain, IoT, Big Data, Machine Learning (ML), among others [3], [4]. Data-driven agriculture through these technologies is the most promising approach to solve current and future issues. If it were possible to generate a large amount of data from farms and use it to drive some agricultural decisions, most of these food problems worldwide could be solved [3]. For example, if it were possible to allow farms to build data sets or maps for diverse environmental factors around the farm, they could implement techniques such as smart farming, precision farming, vertical farming, and others. It has been proven that data-driven agriculture improves crop yields, reduces costs, and guarantees sustainability [5].

Li et al. (2020) explain that Agricultural Big Data is part of cutting-edge technology. It contains concepts, technology, and specific measures covering the entire gamut of agricultural activities, such as farming and planting. The same authors state that by incorporating informatization, intelligence, and precision, the problems of traditional agriculture can be solved. However, the research on Agricultural Big Data is in the initial stage, so more researchers are needed to do more research and analysis [6].

The implementation of Agricultural Big Data involves a set of challenges to consider. From the technical point of view in the implementation, White et al. (2021) explain the challenges in the data, inaccessibility, unusability, incompatibility, inconvenience, lack of data interoperability, lack of rural bandwidth, lack of data calibration, and lack of representation of crop growth models and weather forecasts [2]. Lassoued et al. (2021) analyzed the impact and potential of Big Data in agriculture. The authors identify several challenges, such as data sources and lack of standard, lack of security, cybercrime, and intellectual property protection [7]. Bhat and Huang (2021) point out challenges in information quality, safety, and security [3].

From a societal point of view, Lassoued et al. (2021) identify the lack of staff training to manage large volumes of data [7]. On the other hand, Li et al. (2020) identified challenges from the point of view of the relatively backward rural areas. They warn that there is little understanding from the farmers due to the low level of education in addition to the low amount of online sales, lack of talent to develop Agricultural Big Data in rural areas, and the limited capacity of rural Internet and data processing, rural equipment and facilities cannot meet the conditions of Big Data development. People must rely on agricultural Big Data and thus become a key technology for agricultural development, which can significantly improve agricultural efficiency and reduce the costs of production and sale of products [6].

According to Gopal (2020), due to the multimodal nature of data, it has several challenges, such as improving methods for data collection and selecting effective statistical and data analysis techniques to understand agricultural activities. To improve these aspects, the mechanism used in smart farming is ML, the scientific field that affords a machine the ability to learn without much programming. It has arisen along with Big Data technologies and high-performance computing to create new opportunities to facilitate, quantify and understand the intensive data processes in agricultural operating environments [4].

The developments indicate that agriculture can benefit from ML at every stage, such as species management, field management, crop management, and livestock management [4]. ML is used in a range of agricultural applications including yield prediction algorithms, image recognition algorithms, and robotics to harvest different types of specialty crops [8].

Agricultural Big Data is playing an essential role in the integration of ML. Farmers use the data to calculate crop yields, fertilizer demand, cost savings, and even to identify optimization strategies for future crops [4]. For crops, ML is being used to predict yields, detect diseases, weed detection, crop quality, and recognize species. For livestock ML is being used for animal welfare and livestock production [9].

Due to the volume, variety, and complexity of agricultural data sets, there are many challenges to implementing Agricultural Big Data on farms [10]. The main opportunities and challenges lie in establishing a reference point in the agricultural sector because the factors that affect agriculture will vary with climate, geographic zone, soil type, crop, and traditions [11].

This work aims to discuss the different challenges of using ML in Agricultural Big Data. We want to highlight the technologies used, the kinds of issues that need to be resolved, the ML techniques used, and the challenges imposed by volume, variety, velocity, veracity, and the analysis itself. We provide a framework that summarizes the data to allow researchers to make a better-informed decision on which ML paradigm or solution to use depending on the specific Agricultural Big Data scenario. It also allows identifying research gaps and opportunities in this area. Consequently, this work serves as a comprehensive foundation and facilitator for future research. To this end, we conducted a Systematic Literature Review (SLR), applying the PRISMA protocol [12]. We selected a set of 30 articles that explain the use of Big Data and ML in agriculture.

The structure of the paper is as follows. Section 2 contains the theoretical background on Big Data, Agricultural Big Data, ML, and the main challenges reported in the literature.

 doi:10.20944/preprints202202.0345.v1

Section 3 describes the methodology used to collect the relevant papers for the study. Section 4 contains the results derived from the analysis of the selected papers. Section 5 discusses the main challenges of applying ML in Agricultural Big Data. Finally, Section 6 presents the conclusions.

2. Background

In this section we explain the basic concepts of ML and Big Data. On the one hand, we explain the use of ML in agriculture, and on the other hand, the use of Big Data and its development in agriculture. Finally, we mention the main challenges in Agricultural Big Data, described in the literature.

2.1. Machine Learning

ML is a field of investigation which focuses formally on the theory, performance and properties of learning systems and algorithms. It is a highly interdisciplinary field, based on different areas like artificial intelligence, optimization theory, information theory, statistics, cognitive science, optimum control and many other scientific, engineering and mathematical disciplines [13]. Due to its implementation in a wide field of application, ML has covered almost all the domains of science, having a great impact on science and society [14]. It has been used in a variety of problems, including recommendation drivers, recognition systems, informatics and data mining, and autonomous control systems [15].

Depending on the nature of the feedback available for a learning system, ML can be classified into three main types: supervised learning, unsupervised learning and reinforced learning. Table 1 summarises the main techniques, showing a comparison of machine learning techniques from different perspectives in data processing. The row "Data processing tasks" in the table indicates the problems that need to be solved, and the crow "Learning algorithms" describes the methods that can be used. With the intention of converting raw data into useful data, a preprocessing effort is required. This usually includes: (a) data cleaning to remove inconsistent or missing elements and noise, (b) data integration, when there are many data sources, and (c) data transformation, such as normalization and discretization [16].

Table 1. Main ML techniques.

Classification Type	Supervised Learning	Unsupervised learning	Reinforcement Learning
Data processing tasks	Estimation Classification Regression	Clustering Prediction	Decision-making
Learning Algorithms	Support vector machine Bayesian networks Neural networks Naïve bayes Hidden Markov model	Dirichlet process mixture model X-means K-means Gaussian mixture model	TD-learning Sarsa learning Q-learning R-learning

Briefly, from the perspective of data processing, supervised learning and unsupervised learning focus mainly on data analysis, while reinforced learning is preferred for decision-making problems.

In general, the goal of ML algorithms is to optimize the performance of a task by exploiting examples or past experience. through the exploitation of examples or past experiences. ML can generate efficient relationships with respect to data inputs and reconstruct a knowledge schema. ML will perform better, the larger the volume of data available [16].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

On the other hand, Deep learning (DL) is a branch of ML that tries to model abstractions with a series of algorithms by using a deep layer with multiple processing layers. DL, which is of great interest in the field of artificial intelligence, has come to the fore in natural language processing and image classification [17]. Figure 1 shows the relationship between AI, ML and DL.

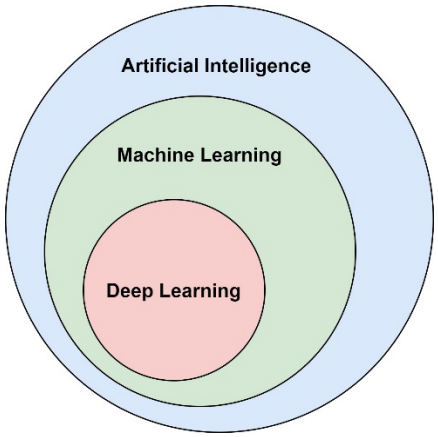

Figure 1. Relationship between AI, ML and DL.

DL it has algorithms such as Convolutional Neural Networks, Recurrent Neural Networks, Restricted Boltzmann Machine, and Deep Belief Network. DL has the advantages of processing unstructured data at the maximum level, producing high quality results, and avoiding unnecessary costs.

ML has been used to solve different problems in agriculture, such as crop management, including yield prediction; disease detection, weed detection, crop quality and species recognition; livestock management including animal welfare and livestock production; water management; soil management [9], [16], [17]. Figure 2 presents a summary of this.

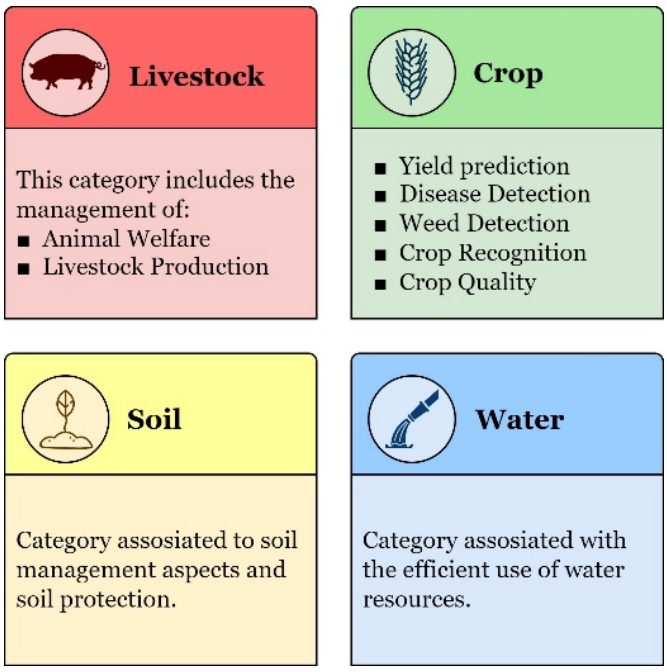

Figure 2. Use of ML in agriculture.

An example of this is that many producers say that weeds are the most serious threat to crop production. Accurate weed detection is very important for sustainable agriculture,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

because weeds are difficult to detect and distinguish from crops. ML algorithms in conjunction with sensors now allow accurate detection and identification of weeds without causing environmental problems or secondary effects. ML for weed detection has led to the development of tools and robots to destroy weeds, minimising the need for herbicides [9]. Accurate detection and classification of the characteristics of crop quality have increased product values and reduced waste. Figure 3 presents a graph showing the different ML techniques that have been used in improving agriculture.

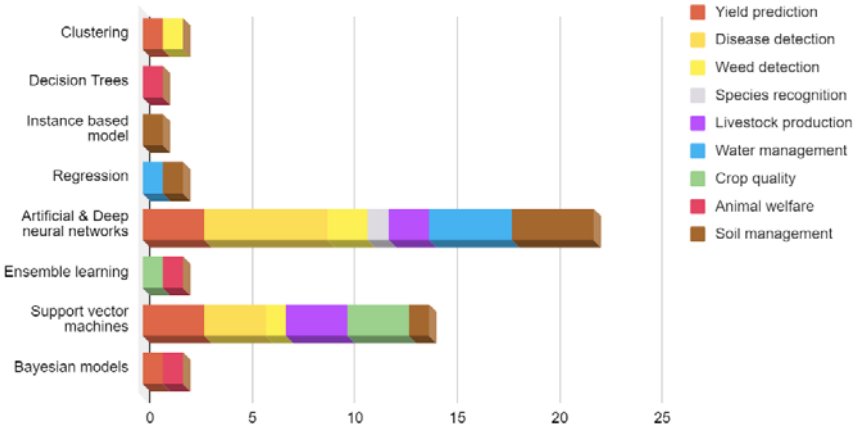

Figure 3. ML techniques used in agriculture [9].

Benos (2021) updates the information in Figure 3 through a systematic study of ML use in agriculture during the years 2018 to 2020 [16]. Figure 4 presents the results obtained.

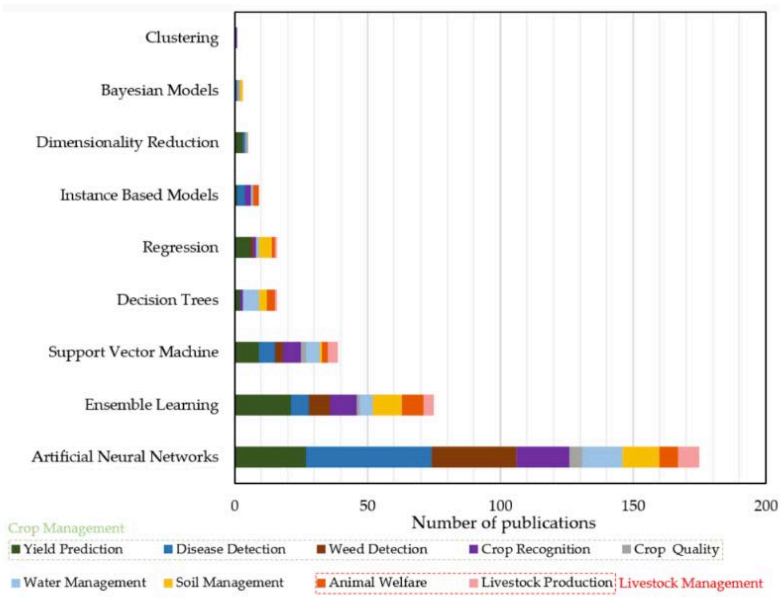

Figure 8. Machine Learning models giving the best output.

Figure 4. ML models giving the best output [16].

It is observed that the ML Artificial Neural Networks (ANN) algorithm is still the preferred algorithm for data analysis. On the other hand, Ensemble learning has been gaining ground, and outperforms the use of other algorithms such as SVM and Decision Trees (DS). According to Benos (2021), the most commonly used data come from meteorology, soil, water and crop quality, remote sensing, satellite imagery, UAVs and UAVs, as well as in situ and laboratory measurements [16].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

The increased interest in ML research in agriculture is a consequence of several factors: the considerable advances in ICT systems in agriculture; the vital need to increase the efficiency of agricultural practices while reducing the environmental burden; and the need for reliable measurements with the handling of large volumes of data [16], [17].

2.2. Big Data

Big Data is defined in four dimensions (4 Vs) [18]. First, it refers to the enormous volume of data that are generated, stored and processed. Second, it also refers to the high velocity of data transmission in interactions, and the rates at which data are generated, collected and exchanged. Thirdly it refers to the variety of data formats and structures (structured, semi-structured and unstructured) which result from the heterogeneity of data sources [19]. The fourth dimension is veracity, which refers to the ability to validate the quality of the data used in the analyses.

Apart from the "4 Vs", another dimension of Big Data must also be considered, namely its value. The value is obtained by analyzing data to extract hidden patterns, trends and knowledge models through the use of algorithms and smart data analysis techniques. Data science methods increase the value of data, giving better understanding of their phenomena and behaviors, optimizing processes and improving the discoveries of machines, business and scientists [20]. We cannot therefore consider the Science of Big Data without including data analysis and ML as key steps for numbering value among the strategies of Big Data Science [21].

In practice, Big Data analysis tools enable data scientists to discover correlations and patterns through the analysis of massive quantities of data from different sources. In recent years the science of Big Data has become an important modern discipline for data analysis [21]. It is considered an amalgam of classic disciplines like statistics, artificial intelligence, mathematics, and informatics with its sub-disciplines including database systems, ML and distributed systems [22].

This is the Big Data Ecosystem that handles the evolution of data, models and support infrastructure throughout its life cycle; it is a whole set of components, or architecture, for storing, processing and visualizing data and delivering results to guide applications [23], [24]. The Framework Architecture of Big Data in Figure 5 includes data storage, information management, data processing, data analysis, and interface and visualization components.

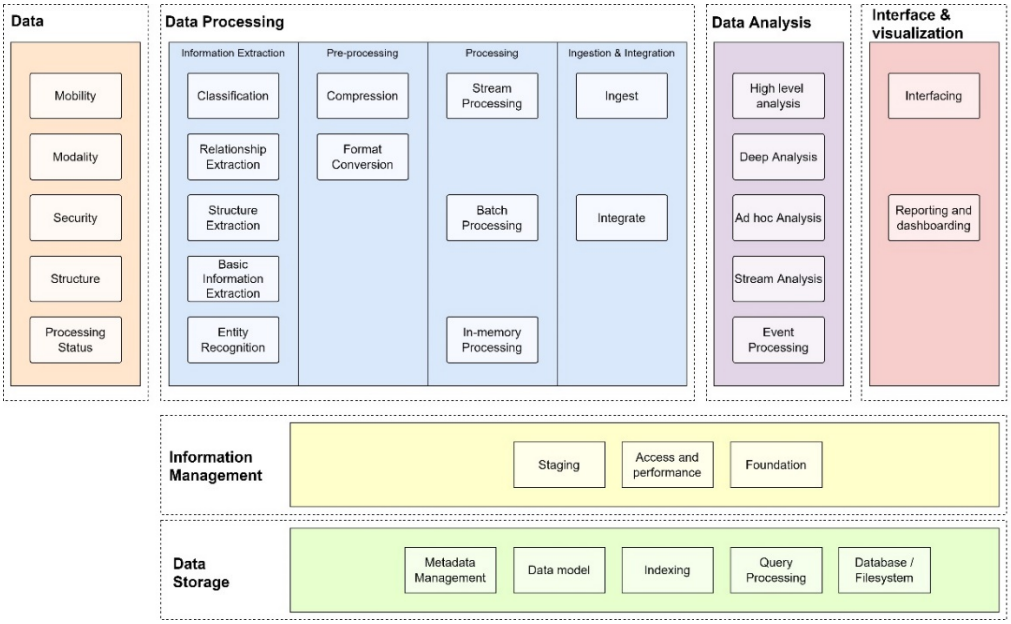

Figure 5. General architecture of Big Data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

As shown in Figure 5, the Big Data process starts with identification of the sources from which useful data are extracted [18]. Next, the data are stored in one of the designed data models, depending on whether the data are structured or not. In the following step, the data are classified and filtered according to the type of analysis required. In the Processing stage, it is defined whether it will be by Bath or Stream, in addition to the memory-based storage [25]. The classified data are analyzed using appropriate tools, for example DL [26], Ad hoc analysis [25], and data science in general [27]. The data obtained must be presented through some kind of visualization tool. Finally the data are analyzed by the decision-makers [24].

Big Data in agriculture refers to all the modern technology available combined with data analysis as a basis for taking decisions based only on data [28]. The following typology will help us to understand the Big data evolution (see Figure 6).

Big Data has been used to improve various aspects of agriculture, such as knowledge about weather and climate change, land, animal research, crops, soil, weeds, food availability and security, biodiversity, farmers' decision-making, farmers' insurance and finance, and remote sensing [29]. It is also used to create platforms which allow the actors of the supply chain access to high quality products and processes; tools to improve yields and predict demand; and advice and guidance to farmers based on the response capacity of their crops to fertilizers, leading to better fertilizer use. Furthermore it has led to the introduction of plant-scanning equipment to follow up deliveries and to allow retailers to monitor consumer purchases, improving product traceability throughout the supply chain [30].

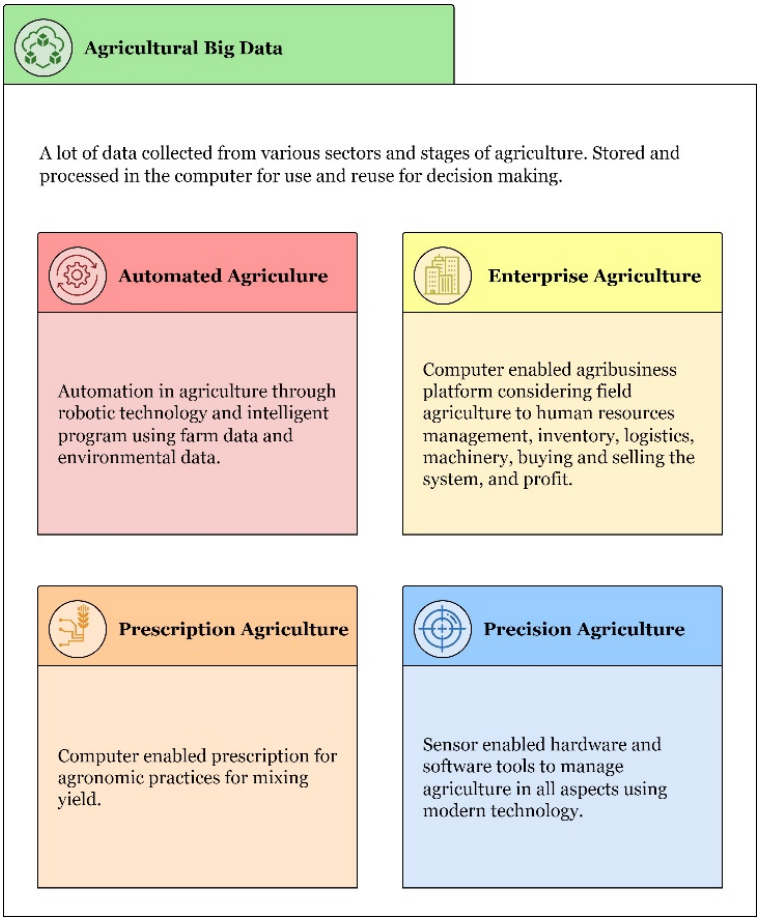

Figure 6. Topologies in digital agriculture.

 doi:10.20944/preprints202202.0345.v1

Big Data does not function in isolation. It has been used in conjunction with other technologies like ML, cloud-based platforms, image processing, modelling and simulation, statistical analysis, NDVI vegetation indices and geographic information systems (GIS) [29]. ML tools have been used in prediction, grouping and classification problems; while image processing has been used when the data are extracted from images (i.e. cameras and remote sensing) [29].

2.3. Challenges in Agricultural Big Data and ML

Several authors explain a number of challenges when using Big Data or ML in data analysis for agricultural development.

White et al. (2021) conducted a survey with researchers participating in a conference on Precision Farming to identify the challenges in different scenarios where Agricultural Big Data is used: (1) mid-season yield prediction for real-time decision-making, (2) sow lameness, (3) irrigation in cotton management, (4) in-season decision making, (5) policy-maker perspective, (6) cropping selection system, (7) business analytics for agriculture, (8) grower perspective, (9) consumer perspective, (10) benchmarking scenario—comparing individual grower yields to modeled outputs based on other people's data [2]. The challenges indicated in these scenarios are: error in the data, inaccessibility, unusability, incompatibility, and inconvenience. An example of this is the lack of data interoperability that prevents integration and unified analysis of data collected by multiple sensors and platforms. Another example is the lack of rural bandwidth that often makes data transmission, particularly in large data sets that include images, impossible. In addition, sensor data need calibration. Finally, they indicate that better representations of crop growth models are required as well as more specific weather forecasts for individual farms and fields [2].

Lassoued et al. (2021) analyzed the impact and potential of Big Data in agriculture. They identify several challenges, such as data sources, given that not all the segments in the value chain capture data the same way. They also point out that there is no standard by which the data are captured, which makes it difficult to harmonize and compile data from various sources [7]. They note through a survey that the implementation of Big Data in an organization depends on a clear strategy and the means to execute it as well as trained personnel to administer large volumes of data. Training and talent, more than capital, are fundamental to optimal production in the future [7]. Another major obstacle identified is data governance. Although most of the experts surveyed indicated their willingness to share their own data under certain conditions, many expressed concerns about data privacy, security, cybercrime and the protection of intellectual property.

Bhat and Huang (2021) conducted a study on the application of artificial intelligence and Big Data in agriculture. They indicate several challenges when applying Big Data in real life. One of these challenges is the compilation and analysis of large volumes of data produced through IoT networks and wireless sensor networks, which include digital images and more data from UAV, satellites and data integration, and pose difficulties for the effective execution of smart farming. The authors explain that most Big Data systems are adequate for large industrial farms because they have the infrastructure to access data, resources and, most importantly, funding. Yet they found few examples in small farming operations in the developing world. Big Data has the potential to support non-industrial farms; however, the moral and ethical questions with respect to availability, cost and financing must be addressed to achieve these advantages [3].

On the other hand, Bhat and Huang (2021) examine challenges in data collection and analysis. The combination of data from a variety of sources causes concern about the quality of the information and its merging, as well as the access to the volume of information compiled causing concern about security and protection. The compiled data sets are enormous and complex, which makes it difficult to manage the normal procedures of smart analysis. These methods do not normally work well when applied to agricultural data.

 doi:10.20944/preprints202202.0345.v1

The authors expect that scalable and versatile methods will adapt to the large amounts of information [3].

According to Gopal (2020), the great challenges of Agricultural Big Data exist at different stages, like data collection, storage and analysis, since the agricultural data set contains various data such as soil, climate, seeds, cultivation practices, irrigation facilities, fertilizers, pesticides, weeds, harvesting, post-harvest techniques, and others. The data are generated and maintained by governments, universities, research organizations, farming companies and agricultural input companies for agricultural production, insurance, marketing, supply chain, packaging, distribution, etc. [4]. Due to the multimodal nature of the data, there are several challenges such as improvement of the methods for data collection, statistics techniques and effective and efficient data analytics to understand and support the functions of several agricultural verticals. On the other hand, Weersink (2018) explains that the data must be collected consistently and fulfill the protocols that can group them into centralized servers which must be protected from cyberattacks while they mask the identity of the individual operation managers [31].

Our paper analyzes the main challenges in Agricultural Big Data when incorporating ML for data analysis. The challenges are classified from the point of view of the intrinsic characteristics of Big Data, the 4 Vs, and in the data analysis itself with ML. From the data found, we propose a framework that summarizes the challenges, the ML techniques and the main technologies used, in order to provide information for future research.

3. Methodology

The research method applied in this article is a SLR. For the selection of articles, the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) method was applied, which contains four stages: identification, screening, eligibility, and included [12]. The databases selected for the search stage are: Scopus, Springer, ACM, IEEE, MDPI, and Web of Science. The search string used in all data sources was "Big Data" AND "Machine Learning" AND (agriculture OR farm). These keywords must be contained in the title, or in the keywords of the article, or in the abstract of the articles. In addition, only scientific articles and conferences in English published from 2015 to 2020 were examined.

In descending order, 580 potential articles were identified in the Scopus database, 567 in the Web of Science database, 486 in Springer, 356 in IEEE, 309 in ACM, and 270 in MDPI. During this process, books, book chapters, working papers, and press articles were excluded. This resulted in 30 relevant articles.

In the identification phase, 2568 articles were identified. This was followed by a screening when the duplication criterion was applied. This resulted in 1489 articles identified. The abstracts of these articles were then reviewed and checked whether or not they dealt with the nexus between agriculture, Big Data and ML. After eliminating the non-relevant articles, 54 articles were saved for detailed analysis. The detailed analysis consists of reviewing the entire article, to ensure that it includes a description about the Big Data process applied and the ML techniques used. Most of the excluded articles dealt with purely theoretical, technological, or experimental issues. Some of the excluded articles dealt insignificantly with the nexus between agriculture, Big Data, and ML. Finally, the SLR was based on 30 relevant articles. Figure 7 summarizes the steps of the relevant article selection process.

The composition of the articles analyzed in depth is very varied, the 30 articles were published in 12 conferences and 18 journals. The journals that include more than 3 publications are International Journal of Emerging Trends in Engineering Research, and Computers and Electronics in Agriculture. The most frequent year was 2020 (see Figure 8).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

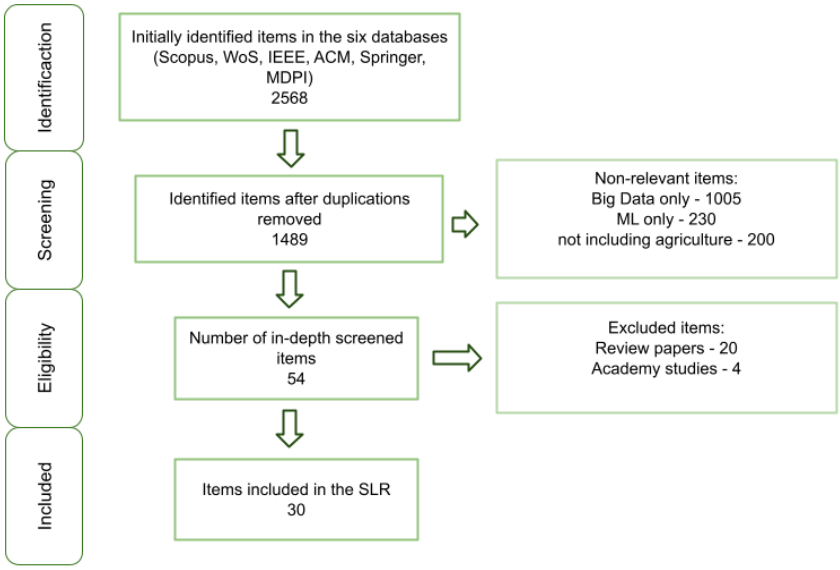

Figure 7. Flowchart of the literature selection process.

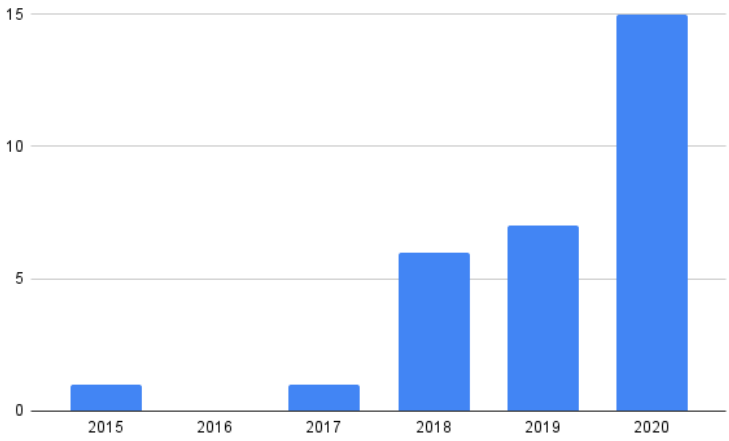

Figure 8. Yearly distribution of the included articles.

4. Results

This section details the main results from the analysis of the selected papers. The analysis stems from a series of Research Questions such as: (1) what kind of problems are solved using Agricultural Big Data and ML, (2) what is the agricultural line of business in which the problems are attempted to be solved, (3) what are the main ML techniques that have been used to analyze the data, (4) what technological tools are used to implement Agricultural Big Data, and (5) what are the challenges to implement ML in Agricultural Big Data.

First, the main solutions described in the context of Agricultural Big Data are explained. Then the use of ML techniques mentioned in the papers is described. The main technologies implemented are also described. Finally, the main challenges described in the selected papers are mentioned.

4.1. Solutions in Agricultural Big Data

The 30 selected articles explain Big Data and ML solutions for problems faced in different areas of agriculture. Solutions were found for Farmer's decision making, Crops, Animal's Research, Land, Food availability and security, Weather and climate change, Weeds. Figure 9 shows the number of papers found by category.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

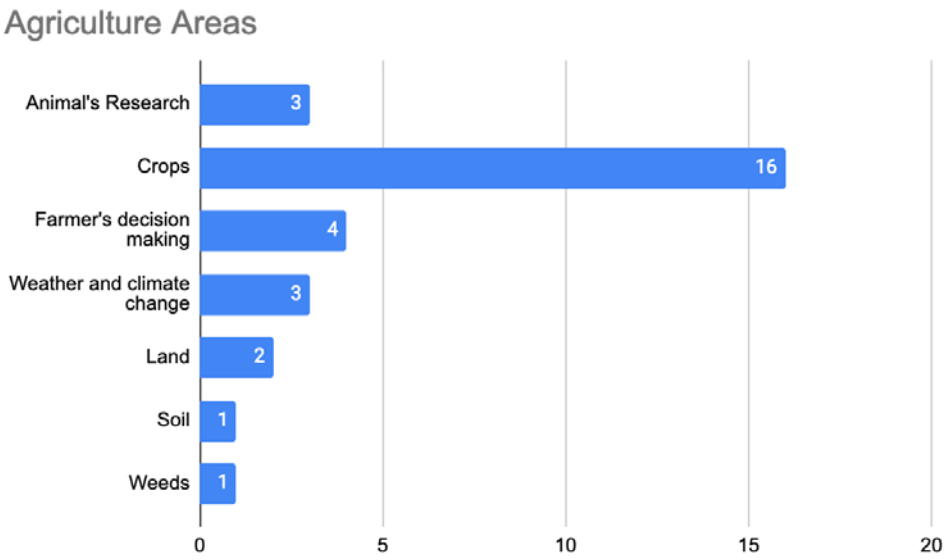

Figure 9. Number of solutions in Agricultural Big Data and ML by industry.

4.1.1. Farmer's decision making

For the case of farmers' decision making, three problems have been noted that the farmers must address, and that have been solved by applying ML in Big Data systems. The first is to lower production costs. Dutta et al. (2015) manage to capture data from domain knowledge on farming processes, understanding of the soil, harvest optimization based on the climate conditions, and data from the farmers' undocumented experiences. With this, they analyze the data to develop low-cost planting [32]. Doshi et al. (2018) developed a Big Data system they call AgroConsultant, which is designed to help Indian farmers make an informed decision about which crop to grow based on the planting season, the geographic location of their operation, the characteristics of the soil and environmental factors like temperature and precipitation [33]. The second problem they address is climate prediction. Rehman (2020) used real-time applications on sensors to capture climate changes in the soil and the atmosphere to establish planting dates [34]. Finally, the third problem refers to increasing production. For this, Tarik (2020) uses methodological data to predict the cereal production rate in an area characterized by an unstable climate [35].

4.1.2. Crops

In the case of crops, the greatest concern is to increase production, after reducing production costs, increasing quality and finally managing diseases. To increase production, Balducci (2018) analyzes environmental data such as climate, humidity, and wind along with production and structural data like type of soil and land extension [36]. Priya (2018, 2020) presents a precision farming model to suggest to farmers what crops must be planted in terms of field conditions [37], [38]. Shelestov (2020) does the same, but using data from satellite images [39]. Ramaraj (2020) analyzed the rice cultivation methods based on the most used varieties of rice, the yield parameters and the morphological characteristics of the crop so as to improve the process to increase production [40]. Yoki (2020) implements the BMS system (Big Data Application Machine Learning-based Smart Farm System) with an emphasis on crop productivity and the importance of increasing the farmers' income. The author concludes that the information and the processable knowledge must be improved at farm level to increase production in addition to improving the quality of the harvest by getting a good price [41]. Kedarmal (2020) uses an ontology of smart farming that contains agriculture-related concepts and properties to link the stored data with a knowledge graph. This graph allows farmers to take measures in time

 doi:10.20944/preprints202202.0345.v1

to improve production [42]. Yahata (2017) used image detection methods obtained from a cyber-physical system to collect data on the stage of crop growth and environmental information [43]. They developed useful rules for an adequate crop adapted through the detection of flower pods.

4.1.3. Animal's Research

In Animal's Research we found only three papers that explain the use of Big Data and ML. In Nobrega (2018) they use an animal behavior monitoring platform based on IoT and cloud computing technologies, in order to monitor sheep inside vineyards. The system allows knowing what each sheep consumes to ensure that the vineyards are not damaged and improve quality. On the other hand, the system allows to keep track of sheep diseases [44]. Abbona (2020) develops a Genetic Programming approach that includes white box techniques that are suitable for the selection of important variables to generate simple models to understand the causes of calf deaths. Ferreira (2019) evaluated beef cattle production performance in Brazil, where the level of animal nutrition is measured to improve quality [45].

4.1.4. Land

In Land, we found only two papers that explain the use of Big Data and ML to solve their problems. Agriculture and Agri-Food Canada (AAFC), is the federal department responsible for agriculture that produces the Annual Crop Inventory in Space maps to improve agricultural production. These maps are valuable operational space-based remote sensing products that cover agricultural land use and non-agricultural land cover found within Canada's agricultural acreage [46]. Similar work was done by Amani (2020) to develop high spatial resolution (30 m) reference maps of the cropland extent of South Asia. The author explains that there is a need to improve production due to the food insecurity experienced in the area [46].

4.1.5. Weather and climate change

On the other hand, we found two papers under the heading of Weather and climate change. In Sathiaraj (2018) they analyzed more than 3,000 weather observation sites in the United States, with the objective of classifying the type of weather in regions across the country [47]. The goal was to understand the climate type of a specific region as it has applications in public health, environment, actuarial science, insurance, agriculture, and engineering [47]. On the other hand, Amaechi (2020) uses various ML techniques to analyze climate data and improve prediction [48].

Kaur (2019) indicates that an essential issue for agricultural planning is to estimate evapotranspiration accurately, as it plays a key role in scheduling irrigation water to use it efficiently [49]. They use ML and Big Data techniques to create an H2O model framework to determine daily ETo. In Ryan (2018) they analyze crops using Makov chains, focusing on weed control and management [50].

In general, Big Data and ML are two technologies that are used to solve various problems in the field of agriculture. The problems described include increased production, increased quality, disease control, cost reduction, climate prediction and control, and crop monitoring. Figure 10 presents a summary of the number of papers found. Some papers reported more than one problem to be solved, for example increasing yield and also crop quality. On the other hand, Figure 11 presents a map summarizing the number of papers selected by category vs. problem to be solved.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

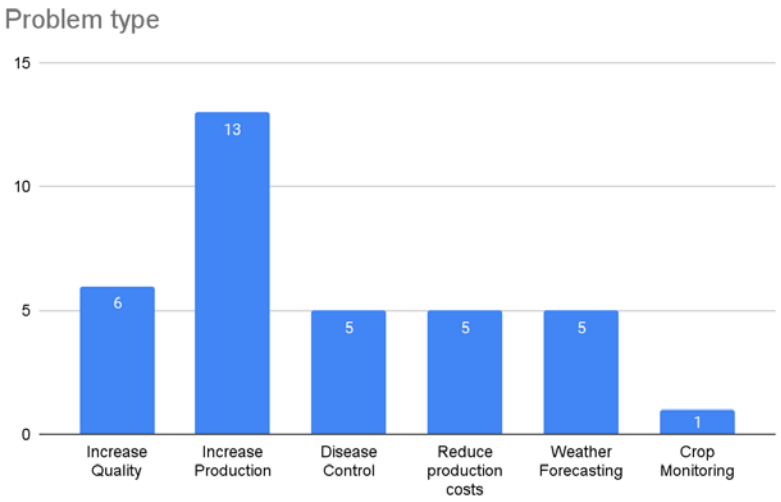

Figure 10. Problems to be solved through Agricultural Big Data and ML.

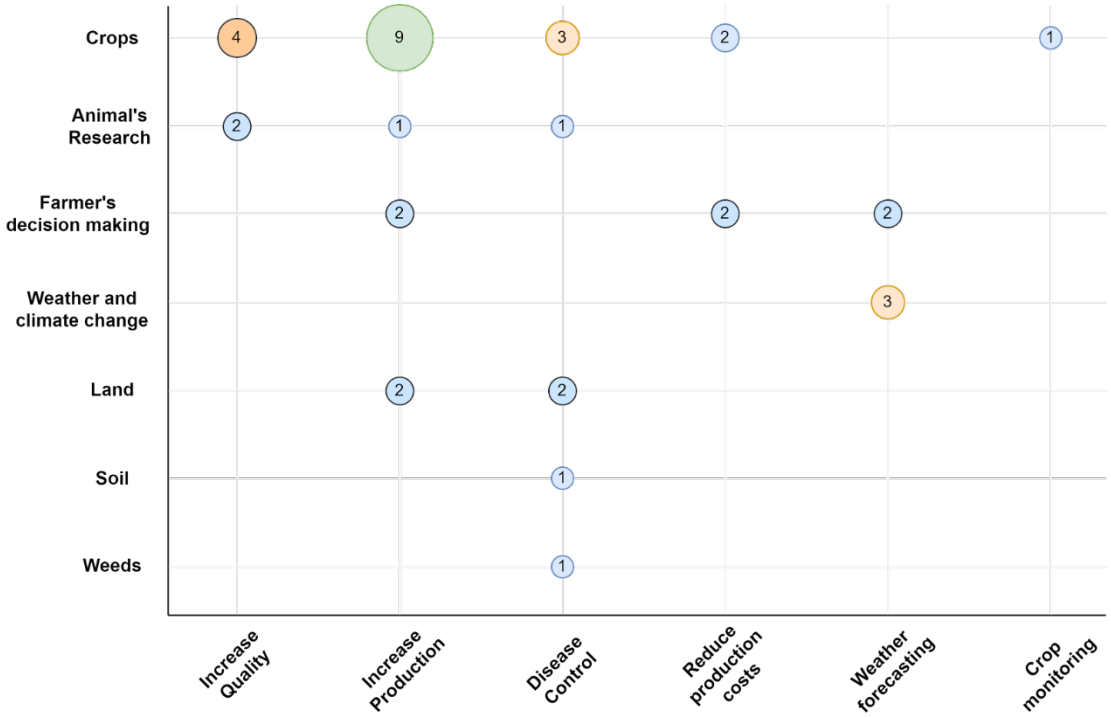

Figure 11. Main problems to be solved vs. agriculture.

4.2. ML Techniques in Agricultural Big Data

From the analysis of the selected papers, a total of 36 different ML techniques were found to be implemented. The techniques were implemented a total of 80 times, as most of the papers implemented more than one ML technique. The techniques that accumulated the most implementations are Neural Networks (NN), Random Forest (RF), SVM and DT. Figure 12 shows the number of implementations found for each ML technique. The uses and characteristics of the most commonly used techniques in Agricultural Big Data and ML systems are detailed below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

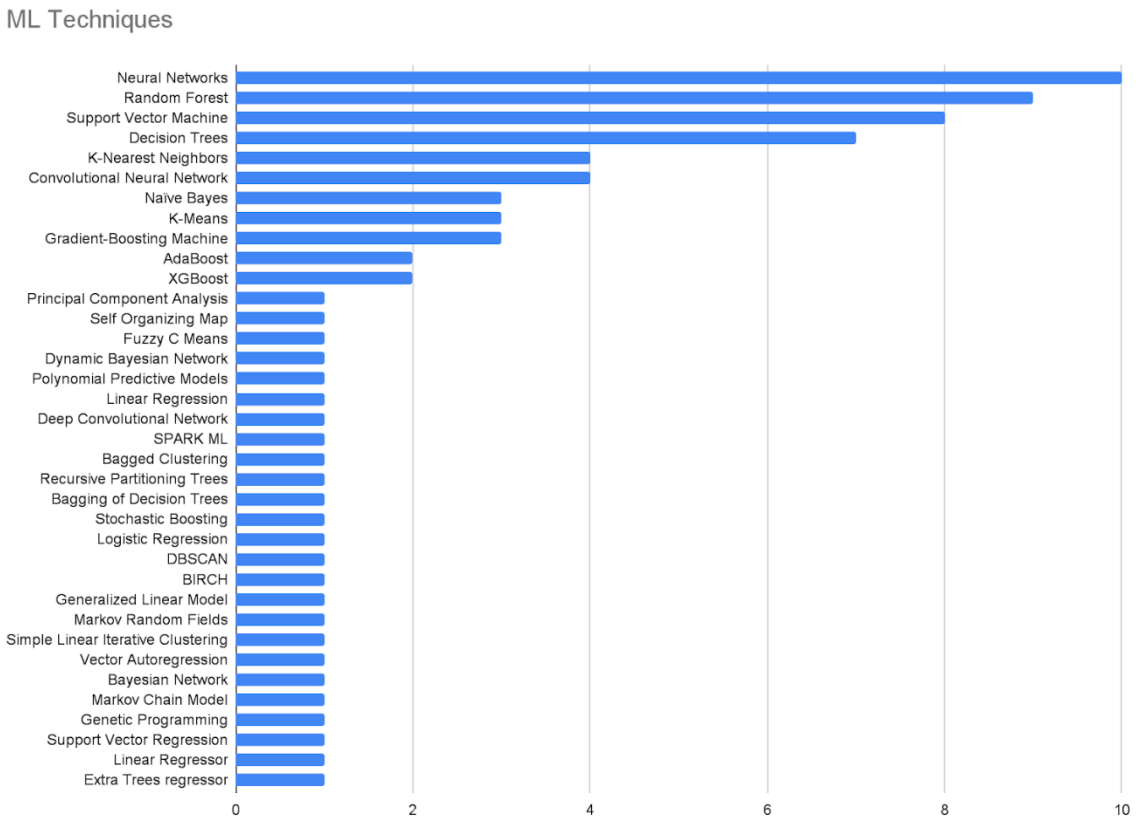

Figure 12. ML techniques used in Agricultural Big Data.

4.2.1. Neural Networks

NN are a good choice for work with big data sets because they have great flexibility to adapt to them, reducing the error produced by adjusting the weights and biases of each neuron based on the data with which it is trained [38]. In Saggi (2019), NN were implemented their performance compared to other ML techniques [49]. NN was the technique with the best performance, avoiding overfitting of the model and demonstrating tremendous abilities for the estimation of daily crop evapotranspiration.

In Doshi (2018), they used NN for the automatic recommendation of crops due to their support incorporated for multilabel classification. The technique performed well in this task, with 91% accuracy in the classification [33]. Shelestov et al. (2020) found that the most sensitive parameters for the accuracy in the classification of a NN are the number of hidden neurons and the alpha regression coefficient. The former had a much greater impact on the overall accuracy of the model than the latter [39]. The authors propose a value of the coefficient alpha with which the best results are obtained, and indicate that the number of hidden layers in the model must be selected independently for each particular case.

According to Priya (2020), NN are efficient at handling data that have no correlation or linearity among them [38], but they are ineffective for modeling time series, which is verified in Balducci (2018), where they implemented a NN and assessed its performance in different tasks, including the reconstruction of missing data in a time series [36]. In this task, the NN implemented presented a considerable prediction error, which, as the authors suggest, may be due to the use of few training data for the one-month time series.

In terms of the architecture of NN, they are structures with a single hidden layer [51], [35], 2 hidden layers [36], [46], [52] and 3 hidden layers [49]. The activation functions used include sigmoid [51], [52], tangent sigmoid [46] and rectified linear unit [49].

 doi:10.20944/preprints202202.0345.v1

4.2.2. Random Forest

According to Priya (2020), some RF applications are harvest prediction, crop yield in adverse conditions, the identification of climatic variables, and the analysis of agriculture-related issues such as nitrogen emissions or drought prediction [38]. RF is ideal for working with massive data sets, since it needs less time to preprocess the data, is competent in terms of global time complexity, and works well with scattered data sets [38].

In Kaur (2019), they verified that this technique behaves efficiently in terms of time complexity when analyzing the computational complexity of the algorithm [49]. Doshi et al. (2018) implement RF for crop recommendation due to the support incorporated for multi-label classification (MLC), and they emphasize that this technique is effective at managing missing values and is resistant to overfitting of the model [33]. This last feature is one of the reasons it is implemented for the classification of farmlands in South Asia in Gumma (2020) [53].

Shelestov (2020) discovered that the maximum depth of each tree and the number of trees in the forest are the most sensitive classifier parameters. Increasing the number of trees improves the accuracy of the prediction; however, it can make the program up to 5 times slower [39]. Sitokonstantinou et al. (2020) used this technique to map rice fields due to the large size of the data set they were working with (satellite images) and due to the ease of the technique to be executed in a distributed manner [54].

4.2.3. Support Vector Machine

SVM is suitable for managing small data sets that do not contain too many outliers, and its performance is increased when the dimensional space of the data is large and the attributes are lower [38].

In Nóbrega (2018), different ML algorithms are compared, among which is SVM, to detect conditions of an animal relative to its position [44]. Of the algorithms analyzed, SVM performed the worst; nevertheless, its results do not differ significantly from the other algorithms and all had a 95% accuracy. A similar case is observed in Yang (2018), where after comparing different ML techniques to predict the growth stage of a plant, SVM was the least accurate technique, although this was over 90% [55]. In both cases, SVM was not the most suitable technique for the tasks carried out, but it did demonstrate a good level of accuracy.

Shelestov et al. (2020) verified that the most sensitive parameters of SVM are gamma, C and the kernel type used [39]. They took measurements on this last point using the kernel's radial basis function (RBF) and sigmoid, and discovered that RBF is the most appropriate for crop classification tasks. Aiken (2019) compared different ML techniques for the classification of pairs of farms [51]. Of these, SVM had the best results in accuracy, sensitivity, specificity and precision. Additionally, it was the technique that required the most runtime, being almost double the others. Nevertheless, the authors conclude that the runtime was not a limiting factor in their study and recommend choosing this algorithm for tasks of the same type.

In Tombe (2020), Fenu (2019) and Vasumathi (2019), SVM was used to determine the health status of the crops, to predict the severity of late blight in the potato, and to predict diseases in fruits, respectively [56], [52], [57].

4.2.4. Decision Tree

The DT is efficient in terms of computation and scalability; moreover, its yield is increased when the same data have no correlation to each other [38]. The efficiency of this technique is confirmed in Nóbrega (2018), where they compared different ML techniques for the classification of the position of an animal using an IoT collar [44]. Of the compared techniques, the authors emphasized DT due to the low computer time required to train the model and the ease of its subsequent interpretation. In addition, it presented one of the best accuracy values and area under the curve (AUC) of the techniques compared.

Another paper where its efficiency was verified is Yang (2018), where the prediction of the growth stage of a plant was studied using different ML techniques [55]. In this paper

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

the authors concluded that the DT was the best compared algorithm due to the low time consumed and high level of precision.

In Balducci (2018), they compared the performance of different ML techniques for the reconstruction of ambiguous or corrupt data captured by IoT sensors, with the DT again being the technique with the best results in most of the experiments [36]. The authors conclude that the performance of this algorithm drops if few data are used to train the model, and that this is affected differently by the different attributes coming from the sensors.

Doshi et al. (2018) compared the DT to other techniques due to its support incorporated for the MLC, for automatic crop recommendation [33]. According to Priya (2020), this technique is not advisable for work with large scattered data sets, since its performance decreases as the data volume increases, and that it does not ideally detect anomalies and manage missing data [38].

4.3. Agricultural Big Data Technologies

The technologies used for the implementation of Big Data and ML systems in agriculture are presented in Figure 13, a cloud of concepts where the size of each concept is directly related to the number of times it was used in the articles analyzed. Of the 30 articles reviewed, the technology that was identified the most times was Apache Spark with a total of 5 Times. Seventy-five percent of the technologies identified were used in only one article and there were 4 articles that did not detail the use of any technology in the implementation of their systems. The use of the most used technologies is detailed below.

Figure 13. Main technologies used in Agricultural Big Data and ML.

For the collection of agricultural data the most common technologies were sensors and satellite imagery. The former were used to take data at the location of interest of: crops, animals, weather or soil properties [41], [38], [44]. Satellite images were mainly used for monitoring land and crops in large areas [46], [53], [54]. These were obtained through satellites or services from external providers such as: Google Earth Engine (GEE), global positioning system (GPS), Sentinel satellites, Landsat satellites or Google Maps.

In the implementation of Big Data systems, the most used file system was Hadoop Distributed File System (HDFS), because it allows separating data sets, storing them in a distributed way in several nodes of a cluster and parallelizing operations on them [54].

 doi:10.20944/preprints202202.0345.v1

Most of the implemented clusters were configured with the various programs provided through the Apache Hadoop framework. Among these the following stand out: Apache Hive and Apache Kafka. Apache Hive was used to configure data warehouses that streamline the process of working with large data sets that were stored in distributed units [58], [59]. Apache Kafka was used for the transmission of information or messages to different nodes of the designed Big Data architecture [59], [41]. The most widely used technology, Apache Spark framework, was mainly used for processing the collected data [44], [58], [34], [59], [54], [42].

The technology most often identified in the implementation of ML models is the Python programming language [36], [53], [52]. Among the articles that use Python for the implementation of the models, most used libraries that facilitate working with the datasets or that implement the models of the ML techniques used in the research. The libraries that were repeated the most are: Scikit-Learn [52], [47], Pandas [47] and NumPy [36].

Regarding the tools for data visualization, Web technologies stand out for their implementation, such as PHP [60] or JavaScript. The latter stands out above all together with libraries such as D3 for data visualization [47], Leaflet.js for displaying maps [33] and React for building interactive user interfaces [47].

In Wang (2019), a Big Data system for agriculture is proposed, designed based on the collection, storage, analysis and application of pear tree data [58]. For the collection of tree growth data (air temperature, soil moisture, light intensity, etc.), a high-precision wireless sensor network is used, whose collected data are sent via TCP protocol to traditional databases (MySQL, MongoDB, etc.). These databases are used temporarily to store the data and serve as data sources for the overall Big Data system. For this purpose, data synchronization software such as NiFi, Sqoop or Flume is used, with which the data sources are synchronized with the HDFS cluster that is responsible for storing all the data together. SparkSQL is used to read, filter and store the data from the HDFS cluster to Apache Hive and Apache Hbase. The former used for data used for analysis and the latter used for data monitoring and visualization of data statistics. Apache Dubbo is used for running farmer management services in a distributed manner. The article does not detail the technologies used for the implementation of ML models.

4.4. Challenges in the use of ML in Agricultural Big Data

Most of the papers selected explain a series of challenges in the use of ML in the Agricultural Big Data system. We describe each challenge according to the intrinsic characteristics of Big Data: volume, variety, velocity, veracity. We also describe the challenges that arise in the analysis process.

Wang et al. (2019) explains that there are challenges in the four stages of Agricultural Big Data: various data sources, low precision, low performance in real time, long collection cycles, high complexity, diversification and lack of appropriate data. There are three main aspects: data cleaning, data consolidation and persistent storage. This is obtained mainly through the use of Hadoop, Hive, HBase and Spark [61].

Priya (2020) concludes that to achieve a good result for the end users, such as farmers and consumers, the data analysis, data summary and methods of data interoperation must be improved at the same time. Nowadays, a computerized approach to agriculture is required to observe and interpret several dangers and treatments such as crop diseases, floods and droughts and to use the resources available more efficiently [38].

4.4.1. Volume

In Yang et al. (2017), a data detection device with sensors and videos is used that communicates with the platform in the cloud with the TCP Socket protocol, loading the data into the cloud in real time [55]. The device transmits data to the platform in the cloud at 3-second intervals. To solve the problem of storing a large amount of data in the system, the authors create a physical division of the table into discrete tables for each day that data are stored. On the other hand, the authors create an analysis service based on Hadoop that

 doi:10.20944/preprints202202.0345.v1

is a file service implemented using a platform in the cloud. The data kept in the MySQL database are transferred to a specific format to be filed in the HDFS system every day. This should be suitable to execute the data analysis service [55]. In Amani (2020) and in Shelestov (2020) they use a large amount of satellite data for downloading [46], [39]. By implementing the work flow on the platform in the cloud, Shelestov overcomes the challenges of downloading and processing Big Data. On the other hand, Gumma (2020) also uses satellite images in the Big Data system [53]. He notes that due to the classification of large areas, the size of the Landsat data is very large (reaching peta-bytes when dealing with a time series), which is why it is very difficult to process the data in regular systems. To solve this challenge, they use the computation platform in the GEE cloud for image processing, because it has the entire Landsat file along with many raster data sets openly available from NASA, the European Space Agency (ESA) and other images that can convey the code to the data. Thus, the complex multi-temporal data on a continental scale can be analyzed using JavaScript or Python simply and can also be shared and replicated by other researchers, reducing the barriers to using supercomputers to perform geospatial analyses [53].

On the other hand, Sitokonstantinou (2019) uses satellite images with a resolution of the 10 m time series, from which it extracts 167 features [54]. The authors report that the automated and prompt acquisition of Sentinel images from the available centers becomes a challenge. Different hubs offer Sentinel data with different specifications, such as their constant archiving policy, data availability, geographic cover and acquisition latency. The authors have developed an application to connect to multiple Sentinel hubs and to seek the pertinent data automatically. This intermediary of Sentinel data recovers the required products from the most efficient center that is decided in terms of download speed and product availability [54]

In Ferreira (2019), a large amount of image data is used to analyze the coincidence of the farms. The authors explain that it is a very challenging task because, generally, the documentation of the attributes of the linked entity (i.e., farm) is highly inconsistent in all the databases due to spelling errors, errors or missing information, and the infeasibility of manual data mining due to the size of the data set [51]. On the other hand, in Saldana (2019) they use input images processed by means of hierarchical convolution modules to reduce the size and gain many more channels [62]. They use from three to five convolutional layers with each image followed by a normalization layer in batches. What they obtain at the end of each module is a grouping layer and an activation layer, then the compressed images are fed into a set of ascending hierarchical sampling modules.

In Abbona (2020), a large volume of data is analyzed to predict whether the calves survive during the 60-day period after birth [45]. They look for a possible solution to highlight the strengths of the young and to find alternatives through variables. The authors indicate that identifying such variables is a complex task but with a solution, since the amount of data recorded among the cattle is enormous nowadays and manageable through ML techniques.

Amaechi (2020) improves climate prediction through a model that uses rules based on the knowledge of experts [48]. The rules have been included in enormous data sets in the data pre-processing.

4.4.2. Variety

In Dutta (2015), they created a database of terrain characteristics based on expert knowledge in the domain, which is generally undocumented information [32]. They used the knowledge of the domain to offer direction to ML. Domain-guided extraction using ML is called semantic extraction, which is why it produced a base of semantic features. The base of meta-features and the base of semantic features were integrated to form a space of enriched features, which was a more significant representation of the heterogenous data.

 doi:10.20944/preprints202202.0345.v1

Tombe (2020) created a crop image characterization scheme that is applied to determine the health status of the crop [56]. Conversely, Rehman (2020) uses heterogenous data such as text, web data and CSV among others, and from this extracts the information needed to construct a set of agriculture rules to provide recommendations [34]. Sitokonstantinou (2019) also uses a set of rules but based on the farmers' expert knowledge [54].

Fenu (2019) implements a module to load data automatically for each crop, a module to load data manually for each crop (used by the farmers during the monitoring survey), and a module to integrate prognosis models [52]. Saldana (2019) manually classified data into six classes of land cover: impermeable surfaces, buildings, low vegetation, trees, automobiles and background [62]. Then, they divided the images into square patches of 256 x 256 x 3 with no overlapping, collecting data from 20,102 images. In Vasumathi (2019) they collect data to fit the quality of the satellite images to analyze plant growth [57]. They achieve contrast stretching or normalization to then perform a noise filtering process. In this process, the pixels in the image show different intensity values instead of real pixel values obtained from the image.

Sathiaraj (2018) uses data from a data set for climatic extreme indices, and data from a National Evaluation of the Climate document published in 2014. The collected data are compared daily with a table of climatic thresholds defined to consider the annual frequency of days that exceeded or fell below a certain threshold. This was carried out for each climate measurement site. This threshold-based data set provides a representation of the climatic extremes at each site and is a resource of useful application to group sites that experience similar trends and climatic extremes [47].

Ryan (2018) uses data from the herbicide resistance testing service at Charles Sturt University, New South Wales, Australia from 2001 to 2015, and data from agricultural surveys applied in several counties registered in the Australian Bureau of Statistics (Australian Bureau of Statistics, 2015) [50]. The first data set consists of annual samples of ryegrass received for herbicide resistance tests on farms all over southern Australia. The locations of the samples were determined by the postal codes of the regions.

In Priya (2020), they use data from such factors as climate conditions, soil type and quality, the variety of the crop and its quality as well as some of the dangerous conditions of plagues, weeds and crop diseases [38]. The author indicates that it is possible to think of agriculture as a significant biological ecosystem, where crops are entities that interact with several bodies in the ecosystem. In addition, none of these entities is fixed and varies dynamically, which implies that to fully understand the agricultural ecosystem, it is necessary to understand the various entities [38]. On the other hand, the author notes that the information compiled is accurate and verified by experts; therefore, when the data are used to develop ML algorithms, the algorithm is more reliable and provides transparency for the development. Once again, the hardware for data collection differs from one organization to another, so the collected data can be in different formats and the results can vary [38]. This is a constant challenge.

4.4.3. Velocity

Of the papers selected, few describe challenges due to the speed with which data must be collected or processed and then visualized. MapReduce is a type of technology used in Big Data to gain greater processing speed with large volumes of data [37]. In Vasumathi (2019), they use the MapReduce algorithm to subdivide the data so that the request is only inspected in the explicit partition, which increased the efficiency and the recovery time of the query [57]. In Saggi (2019), they use MapReduce in the H2O system to obtain more data fragments than the CPU cores [49].

4.4.4. Veracity

In Tarik (2020), they had to perform a pre-processing of the data because the real data were often incomplete (lost values, simplified data), noisy (errors and exceptions) and inconsistent (names, coding) [35]. This caused problems in the implementation of the Big Data system. Conversely, Ferreira (2019) reduced the land space as an alternative in order

 doi:10.20944/preprints202202.0345.v1

to reduce the number of comparisons to be assessed among the attributes of pairs of farms. To verify the data, they had to compare several approaches based on the edition, the Levenshtein and Jaro-Winkler metrics as well as determinist, stochastic and ML approaches to classify the pairs of farms as coincident or noncoincident [51]. The authors point out that they store pairs of similar farms, verifying the values of each approach. In Donzia (2020), the performance of the ML module was improved by using an automatic controller for data stored continuously in HDFS [41].

Saldana (2019) prepared the training data for the first CNN model, dividing the original aerial images in square patches with a predefined resolution [62]. This led them to increase the sample of training data. They used techniques to increase data that had to be verified to improve the data quality. On the other hand, in Vasumathi (2019) they adjusted the quality of the satellite images to analyze plant growth. The adjustment was due to insufficient quality [57]. They carried out a contrast stretching or normalization process and another process of noise filtering of the pixels with different intensity values to obtain real values of the pixels obtained from the image. In Yahata (2017), they use data from photos of flowering plants to count them and examine the growth [43]. The problem is verifying the data when the flower is very close to another one, since the incorrect coincidences must be reduced. The authors use a robot to improve the accuracy of the flower pairing. Sitokonstantinou (2019) developed a framework for updating images of the plots to validate the land cover data through ML algorithms. With the validated data, detailed maps of land cover at plot level were created that can distinguish between rice crops and other types of crops [54].

4.4.4. Analysis with ML

Dutta (2015) investigated how to perform tasks of learning, inference and prediction with Linked Open Data [32]. Experts in the agricultural field and farmers were able to potentially define the complete space of features needed to optimize the harvest. Often, that could be in a casual format or unstructured knowledge that renders it inaccessible from the point of view of the system. The authors used four rule builders to formulate these relations: fuzzy rule builder, conditional probabilistic rule builder, order logic rule builder, and a threshold-based event rule builder (where the threshold of a few environmental variables, defined directly by the farmers together with an event that led to making an unusual decision). Based on the specific rule of the domain or the relationship structure, an ontology translator was created for automatic reasoning from the dynamic time series data from the environmental sensors and the networks of sensors. A task of this translator was to convert the knowledge of the domain into a format that could be used in the functional block called “Semantic signal translator”. Taking this challenge, the system can analyze large volumes of environmental data using ML approaches [32].

Tombe (2020) proposed a computer viewing technique for crop image characterization applied to determine the health status of the crop [56]. With these data a deep convolutional neural network can be used to extract and represent features of the image, and then these features are fed into the support-vector machine for training and subsequent image interpretation. Gumma (2020) employed a similar process with crop images to use multiple decision trees to assign classification labels and to reduce overfitting; in addition, each tree is created from a sub-section of training data [53]. Given that the RF classification is a supervised pixel-based classifier, precise reference data without clouds are needed with high-quality raster input.

Amani (2020) uses satellite images to analyze the type of crops on the plots. The object-based image analysis could improve the classification of the type of crop compared to pixel-based methods. For this, the authors apply the Simple Non-Iterative Clustering (SNIC) algorithm to segment the layered mosaic image [46]. SNIC is an improved version of the Simple Linear Iterative Clustering (SLIC) segmentation algorithm that benefits from a noniterative procedure and applies the connectivity rule from the initial stage. From this

 doi:10.20944/preprints202202.0345.v1

it was possible to use algorithms of deep learning ANN for the classifications of crop types.

Shelestov (2020) needed to process data and configure computer resources for the use of state-of-the-art classification approaches. In order to solve these problems, they developed an automated crop classification work flow based on ML techniques [39].

Wang (2019) uses the decision tree algorithm to analyze pear tree demand. The authors had to predetermine the standard demand for the growth of the pear tree. Each layer of the algorithm is a data comparison process in real time with this standard [58]. The goal is to combine ML with a Big Data platform to extract data features and data values in a short time. The results of this layer are presented to the farm administrator via the data application layer.

In Ferreira (2019), "farm pairing" is done in a scenario with a large amount of livestock data. They compared the performance of twelve automated pairing methods in a different way [51]. They used unsupervised ML approaches (k-meansclustering – KC and baggedclustering– BC) and seven supervised ML approaches (recursive partition trees - RPT, boosted decision trees - BDT, bootstrap classification trees based on BCT, stochastic boosting - SB, support-vector machines SVM, neural networks NN and logistic regression LR). For the probabilistic and ML approaches, they considered both the Levenshtein metric and the Jaro-Winkler similarity criteria. The authors conclude that SVM combined with the Levenshtein metric produced the best results of all the approaches, with almost perfect precision, sensitivity and specificity along with very high accuracy.

Saggi (2019) implemented a multilayer deep learning model, considering multiple hidden layers and a rectified linear active function [49]. The model was trained with stochastic gradient descent by means of back-propagation. On the other hand, Pandya (2020) use series prediction methods such as autoregression (AR), autoregressive integrated moving average (ARIMA), and vector autoregression (VAR), since they faced the challenge of the time interval concept in the transmission data, i.e., the properties of the transmission can change over time [59]. Another challenge was the efficiency of the system to update the ML models based on these algorithms to address the time interval of the concept. The authors proposed a novel framework to address both challenges.

Abbona (2020) used a more frequent set of variables that were encapsulated in the models constructed by Genetic Programming to investigate their zoological meaning in calf production, evaluating the performance of the prediction models [45]. The authors note that the method worked well, which implies that the ML horizon must be investigated further and that comparisons with other techniques must be made, even in larger data sets that contain more features. Evolutionary algorithms can be applied to zootechnical data, obtaining performance models capable of learning the available data.

Amaechi (2020) use a model designed by the authors to improve the convolutional network approach they propose. The authors achieve a high level of weather forecast accuracy compared to the alternative methods tested [48]. Table 2 presents the selected papers and the characteristics of Big Data where they describe challenges. We added a column where they present challenges to carry out the analysis with ML.

Table 2. Challenges in Agricultural Big Data and ML.

Authors	Volume	Variety	Velocity	Veracity	Analysis with ML
Dutta (2015) [32]		x			x
Balducci (2018) [36]				x	
Tombe (2020) [56]		x			x
Priya (2018) [37]			x		
Doshi (2018) [33]				x	
Shelestov (2020) [39]	x				x
Nóbrega (2018) [44]				x	
Amani (2020) [46]	x			x	x
Rehman (2020) [34]		x		x	
Gumma (2020) [53]				x	x

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

Gnanasankaran (2020)[40]					
Tarik (2020) [35]				x	
Wang (2019) [58]	x	x	x	x	x
Fenu (2019) [52]		x			
Aiken (2019) [51]	x			x	x
Ochoa (2019) [62]	x	x		x	
Sathiaraj (2019) [47]		x			
Vasumathi (2019)[57]		x	x	x	
Saggi (2019) [49]			x		x
Ryan (2018) [50]		x			
Yang (2018) [55]	x				
Yahata (2017) [43]				x	
Pandya (2020) [59]					x
Priya (2020) [38]		x			
Abbona (2020) [45]	x				x
Sitokonstantinou (2020)[54]	x			x	x
Donzia (2020) [41]				x	
Choudhary (2020)[42]					x
Amaechi (2020) [48]	x	x			x
Cui (2020)[63]	x	x			

5. Discussion

As we have mentioned in this paper, there are several challenges for the proper use of ML in Agricultural Big Data. These challenges are due to the intrinsic characteristics of Big Data, which are volume, variety, velocity, and veracity. To provide a visual map of the ML techniques and challenges faced, we propose a framework composed of 3 main sections considering these aspects in the main components of Agricultural Big Data. The first section of the framework presents the main challenges. The second section shows the main ML techniques used and the context or problem to be solved. The last section shows the most used technologies. See details in Figure 14.

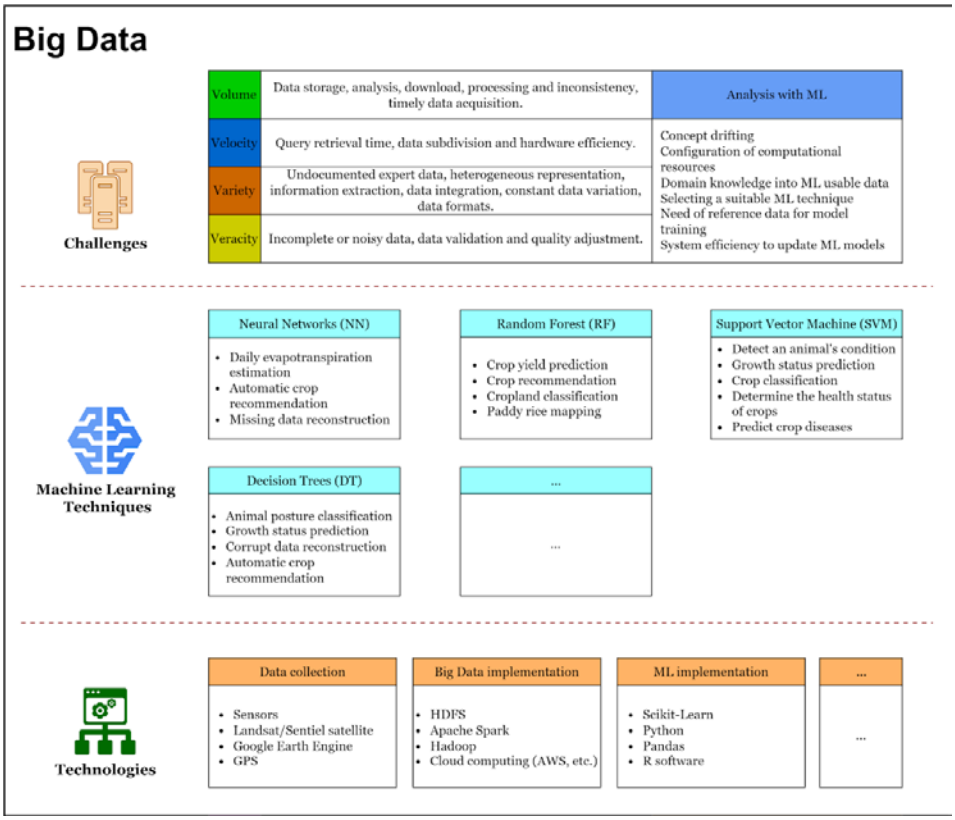

Figure 14. Framework - main ML challenges in Agricultural Big Data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

In this section, we highlight specific challenges relevant to ML in the context of Agricultural Big Data. These challenges are then analyzed from the Vs dimensions. Then, we provide an overview of how emerging approaches address these challenges. Examples of these challenges are unstructured data formats, data input from multiple sources, "noisy" and poor quality data, data scalability, scalability of algorithms, unlabelled data, among others.

As shown in figure 14, the most commonly used ML techniques provide the analyses necessary for predictions, recommendations, situation determination, and automation. These analyses consider techniques like SVR, NN, RF, DT and Naïve Bayes algorithms. A big challenge is to cope with a large volume of data. The SVM algorithm has an O(m3) training time complexity and space complexity of O(m2), where m is the number of training samples. An increase in the value of m will drastically affect the time and memory required to train this algorithm and may even become computationally infeasible for big-size datasets [64]. Another challenge considers the RF algorithm. This algorithm must be tailored for each specific problem to process the data efficiently [65].

Another challenge is the time needed to perform the computations, as this will increase exponentially with increasing data size and may even make the algorithms unusable for large data sets. Other challenges to consider are class imbalance and bias, which will increase as the volume of data increases, causing problems using these algorithms, and may become unable to generalize adequately to new data. These challenges are very relevant as they are present in the most widely used algorithms. Possible solutions to these challenges are using cross-validation and parameter tuning. Figure 15 presents the number of papers per year that use these algorithms. It is observed that NN, RF, and SVM are the preferred ones for the year 2020. On the other hand, Figure 16 represents the relationship of the challenges between ML, the agriculture industry and Big Data, when it needs to be implemented.

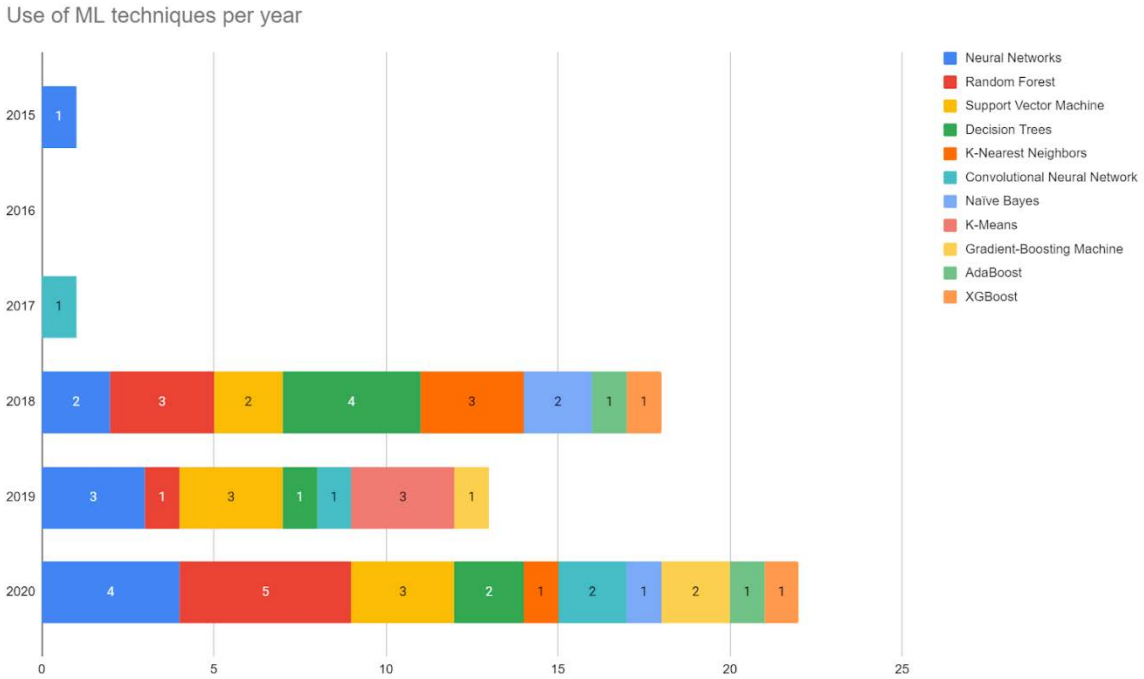

Figure 15. Main ML techniques used in Agricultural Big Data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 February 2022 doi:10.20944/preprints202202.0345.v1

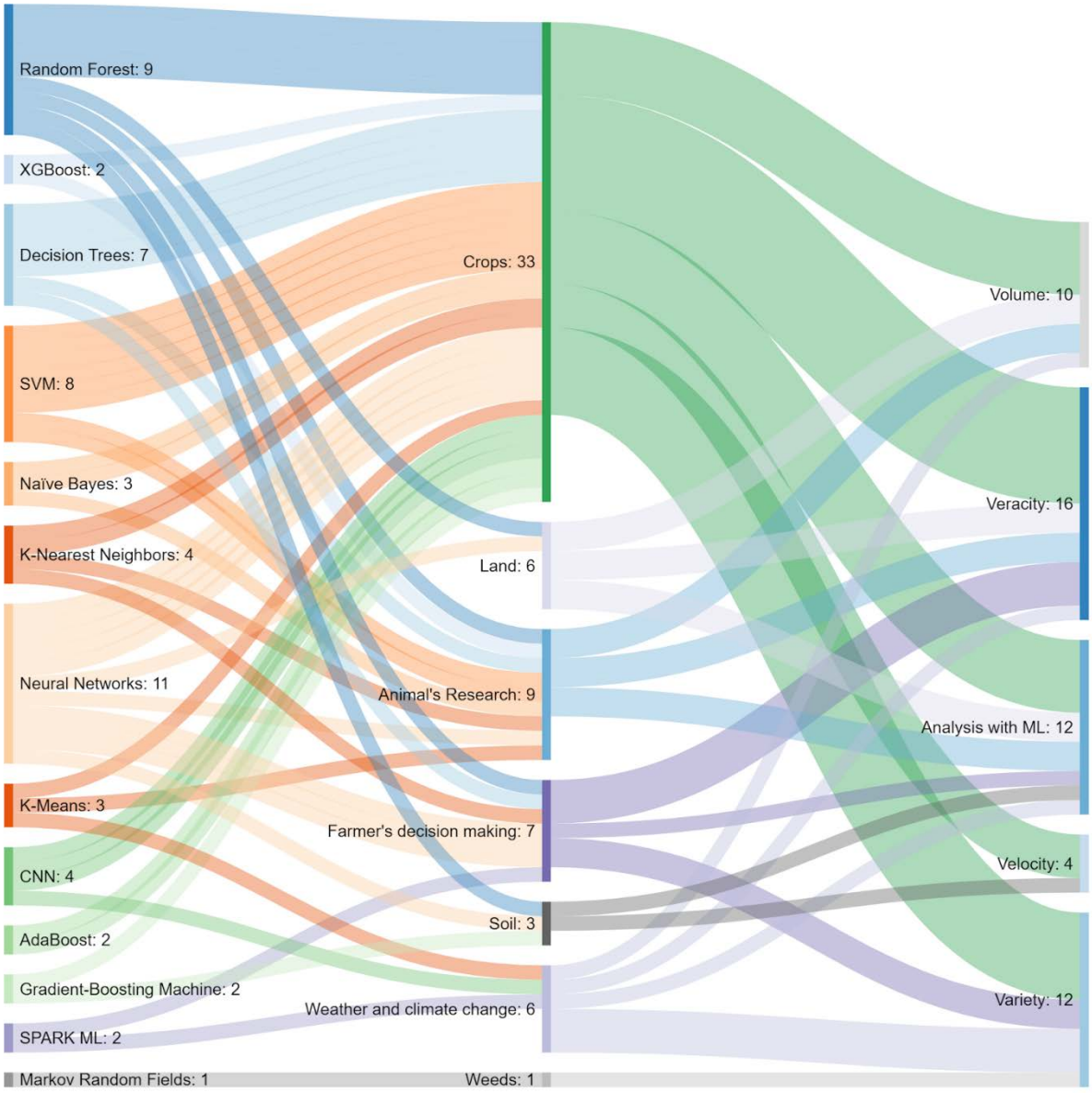

Figure 16. relationship of the challenges between ML, the agriculture industry and Big Data

Another challenge to consider is the variety of data sources (images, videos, sensor data, expert data, among others), as a specific format must be available for the whole data set. This challenge implies that a data Da coming from a data source DSa must have the same format and representation as other Db data from another data source DSb. From a technical point of view, it is possible to manage a heterogeneous database with sufficient metadata to understand the origin, type of data, processing, and changes carried out. A recent technology used to manage the heterogeneous database is the Data Lake [66], [67], which allows managing data in different sections; one for raw data, one for semi-processed data, and one for fully processed data. This technology allows data traceability to be maintained at all times, improving data quality. However, none of the selected papers mention the use of this technology.

On the data processing speed, only two papers mentioned it [57] and [49]. Both authors explain the use of MapReduce to achieve this goal. This aspect is paramount when Big Data is implemented as a solution to a problem because it is a fact that a large volume of data will be used since the data ingestion process in Big Data is constant [68]. On the other hand, it should be considered that Agricultural Big Data is a fully scalable type of

 doi:10.20944/preprints202202.0345.v1

system, which makes the data processing complex, potentially slowing down its processing speed. An example of the above, if we consider a Big Data system that uses data coming only from sensors, then you want to add other sources such as images or videos. From the above, new needs, restrictions and decision-making may appear.

MapReduce has been used to increase data processing speed, also been used to adapt ML algorithms in a different way than the traditional one [69]. Besides, it is possible to split the original data set into subsets and then combine the partial solutions. However, data distribution operations can adversely affect unbalanced data sets. Among the effects is the performance when classifying unbalanced data sets, which may even face the problem of a small sample size, which can be amplified as the original data is distributed on different machines [65]. According to the same author, another effect is the change in the data set that occurs when the partitions of the training and test set are very different between them. Then, it is necessary to design new techniques that can generate synthetic data that best represent minority class instances when using a MapReduce framework [65].

A common assumption of ML is that algorithms can learn better with more data and, consequently, provide more accurate results. However, massive data sets impose several challenges because traditional algorithms were not designed to meet such requirements. For example, several ML algorithms were designed for smaller data sets, assuming that the entire data set can fit in memory [64]. Another assumption is that the whole data set is available for processing at training time. Big Data breaks these assumptions, rendering traditional algorithms unusable or largely impeding their performance [64]. The same author mentions the existence of ML adaptations, such as deep and online learning, to overcome the challenges of ML with Big Data [10]. However, we have identified a budding use of such algorithms in the selected papers.

Regarding the technologies used in Agricultural Big Data and ML, the most used were Hadoop, HDFS, Apache Spark, and Cloud Computing. Hadoop is mainly used to batch process the data and, therefore, the data must be stored in a large repository [70]. Cloud computing is a good solution for storage and processing because satellite data is already available there [46]. Apache Spark is used when streaming data processing is required, and therefore data reading is constant [59].

In Agricultural Big Data, a combination of technologies is required since data from experts, videos and satellite images will be batch processed. On the other hand, data from social networks and sensors will be processed by streaming. For the case of Cloud technologies, there are several tools for the use of ML, Microsoft Azure Machine Learning, which is now part of Cortana Intelligence Suite; Google Cloud Machine Learning Platform; Amazon Machine Learning; and IBM Watson Analytics [71]. These services are offered by established providers, offering not only scalability but also integration with other services and platforms.

From our point of view, the most relevant challenge to consider is the design of Big Data architecture since they must be flexible and highly scalable, considering the architecture design is a complex task [25]. Other challenges are understanding of statistical characteristics of the data before applying algorithms and the ability to work with larger data sets [72]. In addition, specific knowledge is required for certain problems in agriculture such as increased production, quality improvement, and climate change, among others. However, it is important to note that none of the selected papers includes this last aspect as a problem to be faced. According to L'heureux (2017), as data size increases, the performance of algorithms becomes even more dependent on the architecture used [64]. Then, data size increasing makes it necessary to rethink the concept of architecture used to implement and develop algorithms that manage data.

6. Conclusions

This paper presented an SLR using the PRISMA protocol, selecting thirty articles that explain the use of ML in Agricultural Big Data. We analyzed these articles from three

different points of view. First, we recorded the solutions and challenges to different agricultural problems. Second, we reviewed the main ML techniques used in Agricultural Big Data, as well as its main difficulties and challenges. Finally, we recorded the used technologies.

We found 36 ML techniques, considering a total of 80 implementations. Each paper implemented more than one ML technique. The most frequently implemented techniques were NN, RF, SVM, and DT. Despite the positive results described in the papers, the implementation of these algorithms remains a challenge, as there are constant problems due to the increase in data size. These challenges imply adjustments in data classification, the difficult task of calculating the number of training samples, difficulties in the classification of unbalanced datasets, the efficient application of MapReduce due to the increase in data volume, among other aspects to consider.

The most widely used technologies in Agricultural Big Data and ML were Hadoop, HDFS, Apache Spark, and Cloud Computing. Although these technologies can process a large volume of heterogeneous data with reasonable speed, it is still a challenge to use ML algorithms. According to the nature of the data, some adaptations will be necessary to improve the quality of the analyses. A big challenge is the design of Agricultural Big Data architectures since the set of technologies will have to be modified as the volume of data increases.

Author Contributions: A.C. contributed to the planning, organization and direction of the SRL; paper writing; S.P. contributed to paper writing, formatting and creating figures and tables; S.S. contributed to the methodological support and expert judgement; L.M. contributed to the data analysis and discussion. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universidad de La Frontera, Vicerrectoría de Investigación y Postgrado. Dr. Samuel Sepúlveda thanks to research project DIUFRO DI20-0060.

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

References

[1] M. C. Hunter, R. G. Smith, M. E. Schipanski, L. W. Atwood, and D. A. Mortensen, "Agriculture in 2050: Recalibrating Targets for Sustainable Intensification," *Bioscience*, vol. 67, no. 4, pp. 386–391, 2017, doi: 10.1093/biosci/bix010.

[2] E. L. White *et al.*, "Report from the conference,'identifying obstacles to applying big data in agriculture,'" *Precis. Agric.*, vol. 22, no. 1, pp. 306–315, 2021.

[3] N. F. Bhat, S. A., & Huang, "Big Data and AI Revolution in Precision Agriculture: Survey and Challenges," *IEEE Access*, vol. 9, pp. 110209–110222, 2021.

[4] P. S. Maya-Gopal, B. R. Chintala, and others, "Big data challenges and opportunities in agriculture," *Int. J. Agric. Environ. Inf. Syst.*, vol. 11, no. 1, pp. 48–66, 2020.

[5] M. Torky and A. E. Hassanein, "Integrating blockchain and the internet of things in precision agriculture: Analysis, opportunities, and challenges," *Comput. Electron. Agric.*, vol. 178, p. 105476, 2020, doi: https://doi.org/10.1016/j.compag.2020.105476.

[6] L. Hongyan, C. Ziyi, and W. Haitong, "Research of Agricultural Big data," *E3S Web Conf.*, vol. 214, p. 1011, 2020, doi: 10.1051/e3sconf/202021401011.

[7] R. Lassoued, D. M. Macall, S. J. Smyth, P. W. B. Phillips, and H. Hesseln, "Expert Insights on the Impacts of, and Potential for, Agricultural Big Data," *Sustainability*, vol. 13, no. 5, 2021, doi: 10.3390/su13052521.

[8] J. H. Tibbetts, "The Frontiers of Artificial Intelligence," *Bioscience*, vol. 68, no. 1, pp. 5–10, 2018.

[9] K. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine Learning in Agriculture: A Review," *Sensors. Multidiscip. Digit. Publ. Inst.*, vol. 18, no. 8, p. 2674, 2018.

[10] S. Cravero, A., & Sepúlveda, "Use and Adaptations of Machine Learning in Big Data—Applications in Real Cases in Agriculture.," *Electronics.*, vol. 10, no. 5, p. 552, 2021.

[11] H. El Bilali and M. S. Allahyari, "Transition towards sustainability in agriculture and food systems: Role of information and communication technologies," *Inf. Process. Agric.*, vol. 5, no. 4, pp. 456–464, 2018, doi: https://doi.org/10.1016/j.inpa.2018.06.006.

[12] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *PLoS Med.*, vol. 6, no. 7, p. e1000097, 2009.

[13] V. Cherkassky and F. Mulier, "Learning from data: concepts, theory, and methods," *John Wiley Sons, New Jersey*, 2007.

[14] C. Rudin and K. Wagstaff, "Machine learning for science and society.," *Mach Learn.*, vol. 95, no. 1, pp. 1--9, 2014.

[15] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J. Adv. Signal Process. Springer*, vol. 1, no. 67, 2016.

[16] D. Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, "Machine Learning in Agriculture: A Comprehensive Updated Review.," *Sensors*, vol. 21, no. 11, p. 3758, 2021.

[17] B. A. L. Fatih and F. Kayaalp, "Review of machine learning and deep learning models in agriculture," *Int. Adv. Res. Eng. J.*, vol. 5, no. 2, pp. 309–323, 2021.

[18] M. Santos *et al.*, "A big data analytics architecture for industry 4.0," *World Conf. Inf. Syst. Technol. Springer*, pp. 175--184, 2017.

[19] I. SASSI, S. OUAFTOUH, and S. ANTER, "Adaptation of Classical Machine Learning Algorithms to Big Data Context: Problems and Challenges," *2019 1st Int. Conf. Smart Syst. Data Sci. (ICSSD). IEEE*, pp. 1--7, 2019.

[20] D. Gupta and R. Rani, "A study of big data evolution and research challenges," *J. Inf. Sci. SAGE Publ. Sage UK London, England.*, vol. 45, no. 3, pp. 322--340, 2019.

[21] R. Elshawi, S. Sakr, D. Talia, and P. Trunfio, "Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service," *Big data Res. Elsevier*, pp. 1--11, 2018.

[22] B. D. Haig, "Big data science: A philosophy of science perspective," *Am. Psychol. Assoc.*, 2020.

[23] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of Big Data based on its essential features," *Libr. Rev. Emerald Gr. Publ. Ltd.*, vol. 65, no. 3, pp. 122--135, 2016.

[24] Y. Demchenko, C. De-Laat, and P. Membrey, "Defining architecture components of the Big Data Ecosystem," *Collab. Technol. Syst. (CTS), Int. Conf.*, pp. 104--112, 2014.

[25] C. A. Salma, B. Tekinerdogan, and I. N. Athanasiadis, "Chapter 4 - Domain-Driven Design of Big Data Systems Based on a Reference Architecture," *Software Architecture for Big Data and the Cloud*. Morgan Kaufmann, pp. 49–68, 2017, doi: https://doi.org/10.1016/B978-0-12-805467-3.00004-1.

[26] R. Sowmya and K. Suneetha, "Data mining with big data," *2017 11th Int. Conf. Intell. Syst. Control (ISCO). IEEE*, pp. 246--250, 2017.

[27] I.-Y. Song and Y. Zhu, "Big data and data science: what should we teach?," *Expert Syst. Wiley Online Libr.*, vol. 33, no. 4, pp. 364--373, 2016.

[28] M. N. I. Sarker, M. S. Islam, M. A. Ali, M. S. Islam, M. A. Salam, and S. H. Mahmud, "Promoting digital agriculture through big data for sustainable farm management," *Int. J. Innov. Appl. Stud.*, vol. 25, no. 4, pp. 1235--1240, 2019.

[29] A. Kamilaris, A. Kartakoullis, and F. Prenafeta-Boldú, "A review on the practice of big data analysis in agriculture," *Comput. Electron. Agric.*, vol. 143, pp. 23--37, 2017.

[30] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming--a review," *Agric. Syst. Elsevier*, vol. 153, pp. 69--80, 2017.

[31] A. Weersink, E. Fraser, D. Pannell, E. Duncan, and S. Rotz, "Opportunities and Challenges for Big Data in Agricultural and Environmental Analysis," *Annu. Rev. Resour. Econ.*, vol. 10, no. 1, pp. 19–37, 2018, doi: 10.1146/annurev-resource-100516-053654.

[32] R. Dutta, C. Li, D. Smith, A. Das, and J. Aryal, "Big Data Architecture for Environmental Analytics," *Int. Symp. Environ. Softw. Syst. Springer*, pp. 578--588, 2015.

[33] Z. Doshi, S. Nadkarni, R. Agrawal, and N. Shah, "AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms," *2018 Fourth Int. Conf. Comput. Commun. Control Autom. (ICCUBEA). IEEE*, pp. 1--6, 2018.

[34] A. Rehman, J. Liu, L. Keqiu, A. Mateen, and M. Q. Yasin, "Machine learning prediction analysis using IoT for smart farming," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 9, pp. 6482–6487, 2020.

[35] O. J. T. Hajji and Mohammed, "Big Data Analytics and Artificial Intelligence Serving Agriculture," in *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)*, 2020, pp. 57–65.

[36] F. Balducci, D. Impedovo, and G. Pirlo, "Machine learning applications on agricultural datasets for smart farm enhancement," *Machines*, vol. 6, no. 3, p. 38, 2018.

[37] R. Priya, D. Ramesh, and E. Khosla, "Crop Prediction on the Region Belts of India: A Naïve Bayes MapReduce Precision Agricultural Model," *2018 Int. Conf. Adv. Comput. Commun. informatics (ICACCI). IEEE*, pp. 99--104, 2018.

[38] R. Priya and D. Ramesh, "ML based sustainable precision agriculture: A future generation perspective," *Sustain. Comput. Informatics Syst.*, vol. 28, p. 100439, 2020, doi: https://doi.org/10.1016/j.suscom.2020.100439.

[39] A. Shelestov *et al.*, "Cloud Approach to Automated Crop Classification Using Sentinel-1 Imagery," *IEEE Trans. Big Data*, vol. 6, no. 3, pp. 572–582, 2020, doi: 10.1109/TBDATA.2019.2940237.

[40] N. Gnanasankaran and E. Ramaraj, "The effective yield of paddy crop in Sivaganga district – An initiative for smart farming," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 6452–6455, 2020.

[41] S. K. Y. Donzia and H. Kim, "Architecture Design of a Smart Farm System Based on Big Data Appliance Machine Learning," in *2020 20th International Conference on Computational Science and Its Applications (ICCSA)*, 2020, pp. 45–52, doi: 10.1109/ICCSA50381.2020.00019.

[42] N. K. Choudhary, S. S. L. Chukkapalli, S. Mittal, M. Gupta, M. Abdelsalam, and A. Joshi, "YieldPredict: A Crop Yield Prediction Framework for Smart Farms," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 2340–2349, doi: 10.1109/BigData50022.2020.9377832.

[43] S. Yahata *et al.*, "A hybrid machine learning approach to automatic plant phenotyping for smart agriculture," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 1787–1793.

[44] L. Nóbrega, A. Tavares, A. Cardoso, and P. Gonzalves, "Animal monitoring based on IoT technologies," *2018 IoT Vert. Top. Summit Agric. (IOT Tuscany). IEEE*, pp. 1--5, 2018.

[45] F. Abbona, L. Vanneschi, M. Bona, and M. Giacobini, "Towards modelling beef cattle management with Genetic Programming," *Livest. Sci.*, vol. 241, p. 104205, 2020, doi: https://doi.org/10.1016/j.livsci.2020.104205.

[46] M. Amani *et al.*, "Application of Google Earth Engine Cloud Computing Platform, Sentinel Imagery, and Neural Networks for Crop Mapping in Canada," *Remote Sens.*, vol. 12, no. 21, p. 3561, 2020.

[47] D. Sathiaraj, X. Huang, and J. Chen, "Predicting climate types for the Continental United States using unsupervised clustering techniques," *Environmetrics. Wiley Online Libr.*, vol. 30, no. 4, p. e2524, 2019.

[48] E. S. Amaechi and H. Van Pham, "Enhancement of Convolutional Neural Networks Classifier Performance in the Classification of IoT Big Data," in *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, 2020, pp. 25–29, doi: 10.1145/3380688.3380702.

[49] M. K. Saggi and S. Jain, "Reference evapotranspiration estimation and modeling of the Punjab Northern India using deep learning," *Comput. Electron. Agric.*, vol. 156, pp. 387–398, 2019.

[50] I. Ryan, A. Li-Minn, S. K. Phooi, B. JC, and P. JE, "Big data and machine learning for crop protection," *Comput. Electron. Agric.*, vol. 151, pp. 376--383, 2018.

[51] V. C. F. Aiken, J. R. R. Dórea, J. S. Acedo, F. G. de Sousa, F. G. Dias, and G. J. de Magalhães Rosa, "Record linkage for farm-level data analytics: Comparison of deterministic, stochastic and machine learning methods," *Comput. Electron. Agric.*, vol. 163, p. 104857, 2019.

[52] G. Fenu and F. M. Malloci, "An Application of Machine Learning Technique in Forecasting Crop Disease," *Proc. 2019 3rd Int. Conf. Big Data Res.*, no. 76--82, 2019.

[53] M. K. Gumma *et al.*, "Agricultural cropland extent and areas of South Asia derived using Landsat satellite 30-m time-series big-data using random forest machine learning algorithms on the Google Earth Engine cloud," *GIScience \& Remote Sensing. Taylor \& Fr.*, vol. 57, no. 3, pp. 302--322, 2020.

[54] V. Sitokonstantinou, T. Drivas, A. Koukos, I. Papoutsis, and C. Kontoes, "Scalable distributed random forest classification for paddy rice mapping," *zenodo. org*, vol. 11, 2020.

[55] J. Yang, M. Liu, J. Lu, Y. Miao, M. A. Hossain, and M. F. Alhamid, "Botanical internet of things: Toward smart indoor farming by connecting people, plant, data and clouds," *Mob. Networks Appl.*, vol. 23, no. 2, pp. 188–202, 2018.

[56] R. TOMBE, "Computer Vision for Smart Farming and Sustainable Agriculture," *2020 IST-Africa Conf. (IST-Africa). IEEE*, no. 1--8, 2020.

[57] M. T. Vasumathi and M. Kamarasan, "Fruit disease prediction using machine learning over big data," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 556–559, 2019.

[58] X. Wang, K. Yang, and T. Liu, "The Implementation of a Practical Agricultural Big Data System," in *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, 2019, pp. 1955–1959.

[59] A. Pandya, O. Odunsi, C. Liu, A. Cuzzocrea, and J. Wang, "Adaptive and Efficient Streaming Time Series Forecasting with Lambda Architecture and Spark," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 5182–5190, doi: 10.1109/BigData50022.2020.9377947.

[60] A. Shelestov *et al.*, "Cloud Approach to Automated Crop Classification Using Sentinel-1 Imagery," *IEEE Trans. Big Data*, 2019.

[61] L. Zhou, S. Pan, J. Wang, and A. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing. Elsevier*, vol. 237, pp. 350--361, 2017.

[62] K. S. Ochoa and Z. Guo, "A framework for the management of agricultural resources with automated aerial imagery detection," *Comput. Electron. Agric.*, vol. 162, pp. 53–69, 2019.

[63] Z. Cui, X., and Gao, "A Standard Architecture of Agricultural Big Data for Deep Learning. In 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA).," pp. 908–911, 2020.

[64] A. L'heureux, K. Grolinger, H. Elyamany, and M. Capretz, "Machine learning with big data: Challenges and approaches. IEEE," *IEEE Access*, vol. 5, pp. 7776--7797, 2017.

[65] S. del Río, V. López, J. M. Benítez, and F. Herrera, "On the use of MapReduce for imbalanced big data using Random Forest," *Inf. Sci. (Ny).*, vol. 285, pp. 112–137, 2014, doi: https://doi.org/10.1016/j.ins.2014.03.043.

[66] S. M. Wibowo, Merlinda and Sulaiman, Sarina and Shamsuddin, "Machine Learning in Data Lake for Combining Data Silos," *Int. Conf. Data Min. Big Data. Springer*, pp. 294--306, 2017.

[67] B. LaPlante, A., & Sharma, "Architecting data lakes data management architectures for advanced business use cases. O'Reilly Media Inc.," 2016.

[68] Z. S. Khine, Pwint Phyu and Wang, "Data lake: a new ideology in big data era," *ITM web Conf. EDP Sci.*, vol. 17, p. 03025, 2018.

[69] K. Grolinger, M. Hayes, W. A. Higashino, A. L'Heureux, D. Allison, and M. Capretz, "Challenges for MapReduce in Big Data," *Proc. 2014 IEEE World Congr. Serv.*, pp. 182–189, 2014.

[70] J. Loaiza, M. Carmona, G. Giuliani, and G. Fiameni, "Big-Data in Climate Change Models—A Novel Approach with Hadoop MapReduce," *2017 Int. Conf. High Perform. Comput. \& Simul. (HPCS). IEEE*, vol. 45--50, 2017.

[71] C. Yang, Q. Huang, Z. Li, K. Liu, and F. Hu, "Big Data and cloud computing: innovation opportunities and challenges.," *Int. J. Digit. Earth*, vol. 10, no. 1, pp. 13--53, 2017.

[72] S. R. Sukumar, "Machine Learning in the Big Data Era: Are We There Yet?," *Proc. 20th ACM SIGKDD Conf. Knowl. Discov. Data Min. Work. Data Sci. Soc. Good (KDD 2014)*, 2014.