sustainability

MDPI

Review

Data Type and Data Sources for Agricultural Big Data and Machine Learning

Ania Cravero [1,*], Sebastián Pardo [1], Patricio Galeas [1], Julio López Fenner [1] and Mónica Caniupán [2]

[1] Centre of Excellence for Modelling and Scientific Computing, Computer Science and Informatics Department, Universidad de La Frontera, Temuco 4811230, Chile
[2] Information Systems Department, Universidad del Bío-Bío, Concepción 4030000, Chile
* Correspondence: ania.cravero@ufrontera.cl

Abstract: Sustainable agriculture is currently being challenged under climate change scenarios since extreme environmental processes disrupt and diminish global food production. For example, drought-induced increases in plant diseases and rainfall caused a decrease in food production. Machine Learning and Agricultural Big Data are high-performance computing technologies that allow analyzing a large amount of data to understand agricultural production. Machine Learning and Agricultural Big Data are high-performance computing technologies that allow the processing and analysis of large amounts of heterogeneous data for which intelligent IT and high-resolution remote sensing techniques are required. However, the selection of ML algorithms depends on the types of data to be used. Therefore, agricultural scientists need to understand the data and the sources from which they are derived. These data can be structured, such as temperature and humidity data, which are usually numerical (e.g., float); semi-structured, such as those from spreadsheets and information repositories, since these data types are not previously defined and are stored in No-SQL databases; and unstructured, such as those from files such as PDF, TIFF, and satellite images, since they have not been processed and therefore are not stored in any database but in repositories (e.g., Hadoop). This study provides insight into the data types used in Agricultural Big Data along with their main challenges and trends. It analyzes 43 papers selected through the protocol proposed by Kitchenham and Charters and validated with the PRISMA criteria. It was found that the primary data sources are Databases, Sensors, Cameras, GPS, and Remote Sensing, which capture data stored in Platforms such as Hadoop, Cloud Computing, and Google Earth Engine. In the future, Data Lakes will allow for data integration across different platforms, as they provide representation models of other data types and the relationships between them, improving the quality of the data to be integrated.

Keywords: agriculture; big data; machine learning; data type; data source

check for updates

Citation: Cravero, A.; Pardo, S.; Galeas, P.; López Fenner, J.; Caniupán, M. Data Type and Data Sources for Agricultural Big Data and Machine Learning. *Sustainability* **2022**, *14*, 16131. https://doi.org/10.3390/su142316131

Academic Editors: Dionysis Bochtis, Dimitrios Aidonis and Charisios Achillas

Received: 27 September 2022
Accepted: 20 November 2022
Published: 2 December 2022

1. Introduction

Agriculture is an economic activity performed to produce food and is highly dependent on weather and climate. In addition to other activities, agriculture is necessary to sustain human life [1]. Agriculture is affected by climate change in several ways, such as: the increase in extreme weather events, as they impact soil erosion due to droughts and floods; variations in precipitation; changes in average temperatures; changes in pests and diseases; problems with carbon dioxide; changes in the growing season of plants; changes in food quality and nutrition; and changes in sea level [2]. According to Wakelin et al. [3], droughts are expected to significantly impact plant production, as they can impact primary production through effects on plant pathogenic organisms and plant diseases. Some results are the weakening of host defenses or those associated with the rapid evolution of increased pathogenicity within pathogen populations. On the other hand, the impact of climate change on temperature and precipitation is expected to affect various foliar diseases. All these changes affect crop yields, leading to differences in the growth stages of plants during extreme weather events. For these, there is an urgent need to improve agricultural production while minimizing environmental impacts [1].

Digital agriculture (such as agrotechnology and precision agriculture) is a new scientific discipline that promotes agricultural productivity and minimizes environmental impact through data analysis [4]. These data are extracted from current agricultural operations using various sensors, satellite images, and photographs. Data analysis allows for more accurate decision making due to a better understanding of crop dynamics, weather conditions, soil, and the use of farm machinery [4].

As the number of smart machines and sensors on farms increases and a greater variety of data is used, farms will become increasingly data-driven. This is driving the development of Smart Farming through Cloud Computing and the Internet of Things (IoT) [5]. The difference between precision farming and smart farming is that the former was developed for farm management, and the latter considers real-time situations triggered by an event [6]. On the other hand, smart farming includes intelligent assistance in implementing, maintaining, and using IT, allowing farmers to react quickly to sudden changes such as disease alerts or weather events [6,7].

Big Data and Machine Learning (ML) have emerged as two High-Performance Computing (HPC) technologies that use data to create new opportunities for analyzing and understanding complex agricultural processes [4,8]. Some applications belong to two or more domains (Big Data, ML, HPC), such as digital twins or profit simulation engines in the context of Industry 4.0, which include specific hardware and software components optimized for massive workloads [9]. One example is HiBench, a Big Data benchmark offering six categories of workloads, including ML. Another example is BigBench, which covers the collection, storage, and analysis steps of a Big Data system's life cycle. BigDataBench also uses data motifs to generate synthetic data based on real data at scale. These data can be structured, semi-structured, or unstructured, from which the workload is measured to prepare the data for ML testing and execution [10]. On the other hand, significant advances have been acheived in HCP, ML, and Big Data tools. These tools are used in environmental analysis, weather management, weather forecasting, disaster management, water management, and energy management, as well as remote sensing. The increased use of these tools is due to rapid advances in smart technologies, high-resolution remote sensing techniques, and the incorporation of data from social networks [8,11].

A significant challenge for using Agricultural Big Data and ML is the analysis of the large volumes of data produced by data sources and IoT networks. These include data from uncrewed aerial vehicles (UAVs), digital imagery, and satellites, which must be integrated into a large common repository, as most of these data are unstructured [12].

In Big Data Agriculture and ML, structured, semi-structured, and unstructured data are often used, adding complexity to the analysis process as their use is a significant challenge [13]. Unstructured data come from files containing a large amount of information, which is hidden from the data scientist. Examples of such archives are satellite images, surveys, and videos. On the other hand, semi-structured data have been stored in spreadsheets and repositories containing important and unimportant data for the desired analysis. Therefore, it is necessary to process these data to obtain an essential structured data set.

Unfortunately, the processing of unstructured data is not trivial, requiring the use of specialized tools and the knowledge of experts in this area. It also requires selecting the right types of repositories and databases for further processing and analysis [14]. It is, therefore, crucial to identify the available data, the processing necessary, and possible studies based on the generated data, as ML requires test data sets of sufficient quality to achieve the expected learning [15,16].

It is essential to understand the types of data and the sources from which they come, as different types of analysis can be used with ML. According to Nandi et al. [17], analytics can be descriptive, diagnostic, predictive, and prescriptive. The prescriptive analysis is the most complex, as it is responsible for finding a solution among various variants to optimize resources and increase operational efficiency: the more complex the studies to be performed, the more complex the data processing.

According to Firdaus et al. [18], it is essential to know the data type before applying any algorithm. Therefore, the data type plays a vital role in pre-processing and visualization. There are four main data types: numerical, categorical, time series, and text. Numerical data are further classified into continuous and discrete. Categorical data types represent quality; concepts such as "good", "bad", and others are used to define levels. These data must be processed to be represented as numbers and not text.

On the other hand, time series data types occur mainly in logistic recommender systems, when the time sequence is the most important. Text-type data are words that cannot be analyzed as well as numerical values. According to the author, it is important to determine the data to be analyzed with ML as well as whether pre-processing is required to obtain a quality data set. Often, this task is not simple and requires specialists to carry out data engineering processes to transform the data properly [18].

All these aspects must be considered when designing a Big Data system for agriculture, as it requires the interdisciplinary work of several disciplines, such as data scientists, data engineers, computer scientists, agronomists, etc. Furthermore, deciding the configuration of this equipment requires a budget and people's availability to solve the problem. Different technologies are required if structured or raw data are used.

This work aims to identify the data types and data sources used in ML and Agricultural Big Data, whether structured, semi-structured, or unstructured, according to their condition. In addition, we aim to classify the data types according to the data sources found, such as repositories, databases, and platforms, to identify new data types and trends. According to the Kitchemham and Charters protocol [19] and the PRISMA criteria [20], the paper is developed as a Systematic Literature Review.The discussion on data use, challenges, and trends in the development of Agricultural Big Data and ML serves as a basis for future research, as there are still no formal standards and methods for their construction. On the other hand, we provide a knowledge base on the use of data in Agricultural Big Data for agronomists, farmers, agricultural engineers, and others. Our work is essential as we have not found existing literature reviews explaining data types and data sources in Agricultural Big Data. However, there are several reviews on data types in other domains, such as social science [21], smart cities [22], remote sensing [23], recommendation systems [24], and clustering algorithms [16].

The structure of the paper is as follows: In Section 2, the general concepts of Agricultural Big Data, ML, data types, and storage repositories are described. In Section 3 the methodological process is described. Section 4 describes the main results and answers to the research questions presented in Section 3. In Section 5 a discussion of the main challenges and trends is presented. Section 6 presents a description of four threats that challenge the validity or our study. We close the article with a summary of our findings and conclusions.

2. Background

This section includes the following subsections. First, the basic concepts of Agricultural Big Data are described. Next, the following subsection identifies the main ML algorithms used in Agricultural Big Data. This is followed by the basic concepts of data types and storage repositories.

2.1. Agricultural Big Data

Big Data is a type of technology used when the solution is not trivial due to the complexity of the data. It is usually defined through the four dimensions (or 4 V's). The first V represents the volume of data generated from a data source, stored, and processed for further analysis. The second V refers to the variety of data due to multiple structures, structures, and sizes. Data can be extracted as raw or unstructured, semi-structured, or structured data. The third V refers to the speed of data transmission needed for data to be processed and analyzed. The fourth V refers to veracity, which refers to the capability to validate the grade of the data [25].

Big Data allows scientists and engineers to find patterns and trends by examining large amounts of data from multiple origins. A few years ago, Big Data science became an essential modern discipline for data analysis [26]. Big Data includes a mix of classical domains of artificial intelligence, including ML, such as statistics, mathematics, and computer science. In general, it has database systems, ML, and distributed systems [27].

The Big Data process begins with specifying the sources to extract the data required [28]. The next step is storing the data in one of the created representatives, which depends on its processing level, whether unstructured data, semi-structured, or structured. The data are then transformed through filtering and sorting to improve the data quality for various analyses [29]. The next step is to analyze the classified data using analytical tools and algorithms (e.g., Deep Learning (DL), ML, OLAP) [30], as well as general data science [29,31]. This allows decision makers to analyze the data to visualize trends [32].

For example, Semlali et al. [33] use Big Data tools to monitor atmospheric composition. The system architecture contains the data source layer, ingestion, storage via Hadoop, the data management layer, infrastructure, and the monitoring and security layer. They used data on pollutant gas emissions from other sources, such as agriculture, enterprise, and transport. The authors were able to continuously monitor the atmospheric composition through remote sensing. Figure 1 depicts the complete process.

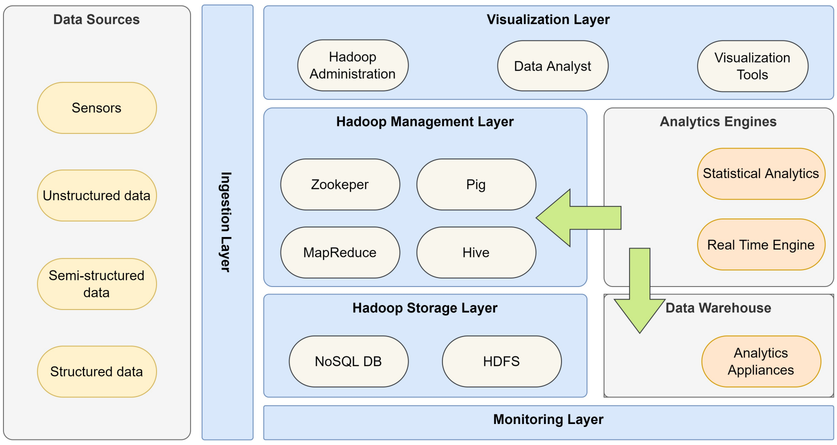


Figure 1. Architecture of Big Data for atmospheric composition monitoring.

MySQL is a database that stores data that have been extracted, processed, and transformed. Then, Python, Java, and BASH are used scripts to read the raw data in Hadoop. Figure 2 shows an example of the process.

Another example is Alex et al. [34], who develop a Big Data system that predicts whether fertilizers will cause disease in crops. They use data such as soil moisture, average rainfall, and soil nutrients. The authors also used data such as phosphorus (P), nitrogen (N), magnesium (Mg), calcium (Ca), and sulfur (S). The Big Data process starts with data enrichment, followed by data clustering, so the data can be classified and analyzed to deliver recommendations. Finally, the Hadoop ecosystem was used to store and process the data analyzed with ML. Figure 3 depicts the complete process.

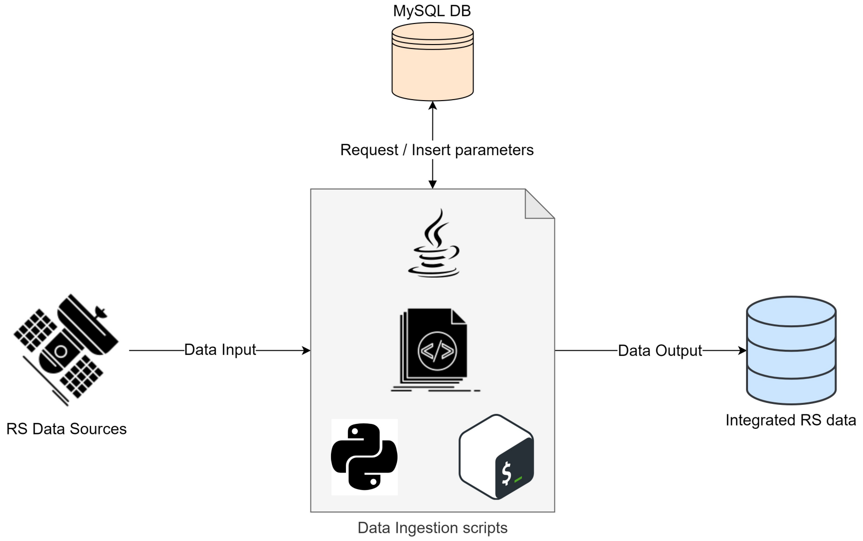


Figure 2. Representation of the ingestion process.

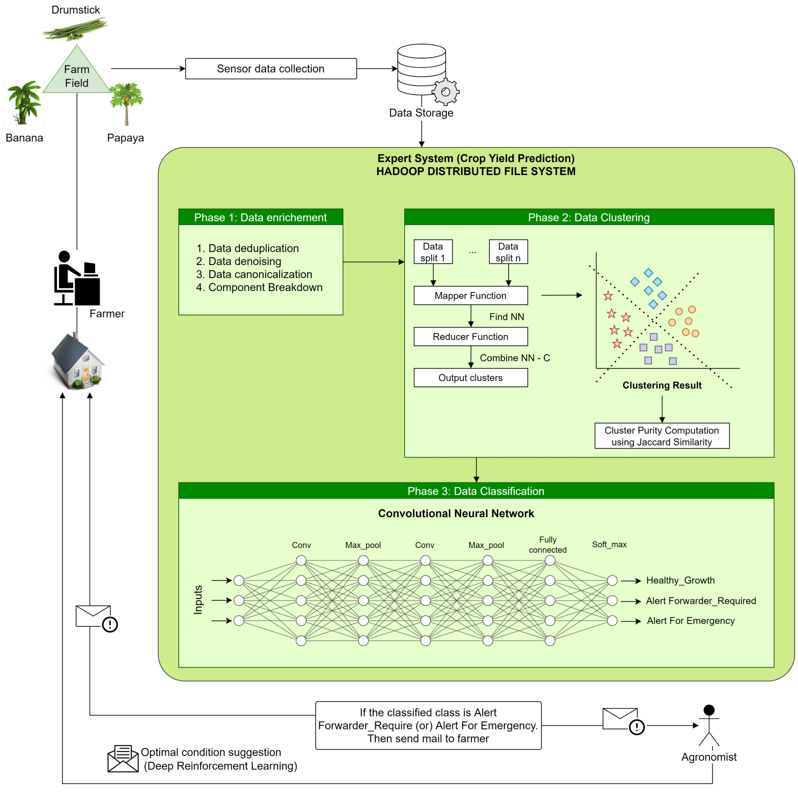


Figure 3. Big Data architecture for fertilizer management and yield prediction.

Big Data enables data scientists and farmers to understand farming behavior, such as weather, land, soil, crops, animal production, weeds, food safety, biodiversity, remote sensing, farmer decision making, insurance, financing, and climate change [12]. It also enables the development of supply chain platforms, allowing agents to access high-quality products, processes, and tools that are capable of improving performance, predicting demand, and addressing farmers according to crop needs, such as the appropriate use of fertilizers.

2.2. Machine Learning

ML is considered an area of research focusing on mathematical theory, system characteristics, and the product of learning algorithms. The investigation process is interdisciplinary, encompassing disciplines such as artificial intelligence, knowledge theory, optimization, statistics, cognitive science, control, mathematics, and engineering [35]. ML is attractive because it can be used in various science domains, significantly impacting society. For example, ML can be used to solve problems such as pattern recognition, recommendation controllers, fact prediction, data mining, and automatic control systems [36,37].

ML can be classified into three algorithms depending on the available data: supervised, unsupervised, and reinforcement learning. Table 1 summarizes these techniques, differentiating them from various points of view, particularly in data processing. The "Learning Algorithms" row explains the methods used, and the "Data Processing Tasks" row contains the problems to be solved. In ML, the data must be structured, so it must be processed in most cases. This task consists of cleaning the data to remove noise and inconsistencies, integrating it if it is drawn from multiple sources, and transforming the data to normalize it [38].

Table 1. Machine Learning techniques.

Classification	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Data processing tasks	Classification Estimation Regression	Prediction Clustering	Decision-making
Learning algorithms	Support vector machine Hidden Markov model Naive Bayes Neural networks Bayesian networks	Gaussian mixture model K-means X-means	Q-learning Sarsa learning TD-learning R-learning

Supervised and unsupervised learning is primarily focused on data analysis. On the other hand, reinforced learning is used for decision-making situations. The ML algorithms presented in Table 1 can optimize the implementation of a task by analyzing samples of data or backgrounds. An important aspect is that ML will function better with more extensive volumes of data to be explored [38].

ML algorithms have been used to improve livestock welfare; increase livestock production; improve yield prediction, crop management, disease detection, weed detection, crop quality improvement, and species distinction; and improve water and soil management [4,12,38,39].

There are numerous challenges when executing ML in Agricultural Big Data since some techniques must be adapted when there is a large volume of data or the data are variable. There are also challenges in validating the data and obtaining a quality set. Solving these challenges is not a trivial task, but proposals have been carried out, allowing progress in this area of research [12].

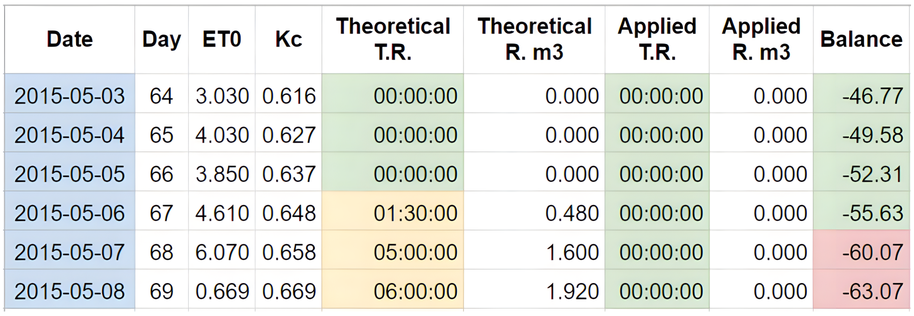
2.3. Data

Agricultural Data refer to variables and/or attributes that farmers require for carrying out their business activities. The data include specific agricultural records or parameters, such as crop varieties, yields, soils in use, extensions, etc., but also business-related information, such as products, suppliers, clients, payments, etc. They are classified into structured, semi-structured, and unstructured data, depending on the storage format in which the data are found [40].

Villars et al. [41] explain that structured data are blocks of information, and unstructured data are raw files. They determined that 23.7% of the data in an organization are structured, while 61.8% represent unstructured data, with a difference of almost 70EB (exabyte, or billion gigabytes). Eberendu et al. [42] explain that 80 percent of an organization's data is unstructured. Examples of these data are those coming from social networks, customer calls, emails, online comments, and information from embedded devices, among others.

2.3.1. Structured Data

Structured data have an established format and, therefore, a certain length in memory; they are also easy to store since they have a logical structure previously defined by the database engine. The data are organized in an identifiable structure, which is easy for people in an organization to access and retrieve [42]. Figure 4 presents an example of structured data, as the data can be stored in a database with a Date-type format. On the other hand, the day is stored as an Integer data type. Moreover, the daily potential evapotranspiration (ET0), crop coefficient (Kc), applied R M3, and Balance are stored as Float-type data since a higher precision is required. Finally, the theoretical RT and applied RT data are Time data types.



Date	Day	ET0	Kc	Theoretical T.R.	Theoretical R. m3	Applied T.R.	Applied R. m3	Balance
2015-05-03	64	3.030	0.616	00:00:00	0.000	00:00:00	0.000	-46.77
2015-05-04	65	4.030	0.627	00:00:00	0.000	00:00:00	0.000	-49.58
2015-05-05	66	3.850	0.637	00:00:00	0.000	00:00:00	0.000	-52.31
2015-05-06	67	4.610	0.648	01:30:00	0.480	00:00:00	0.000	-55.63
2015-05-07	68	6.070	0.658	05:00:00	1.600	00:00:00	0.000	-60.07
2015-05-08	69	0.669	0.669	06:00:00	1.920	00:00:00	0.000	-63.07

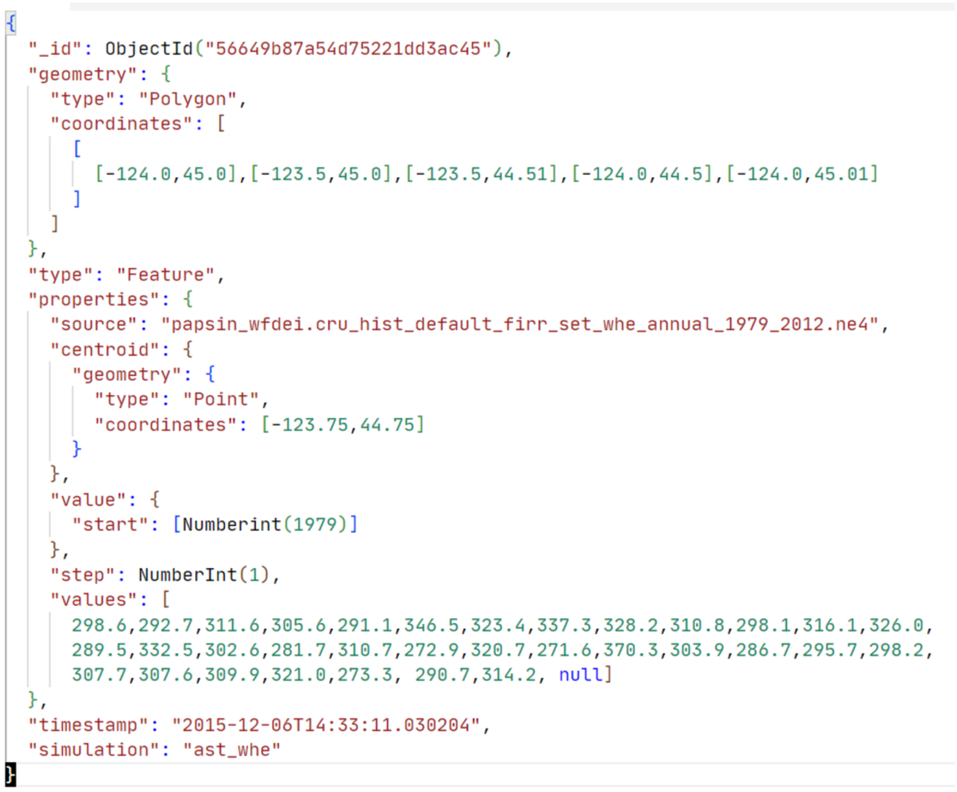
Figure 4. Structured data for irrigation scheduling.

The corresponding data in Figure 4 were stored in a database after applying the irrigation model to a crop, organized as a table. Each row corresponds to a model run. In this way, different columns can be observed where the date of execution of the model, the day of the crop phases, and variables of the irrigation model such as Kc, ET0, water balance, etc., are indicated. Furthermore, the columns with a red border indicate in hours both the theoretical irrigation time and the irrigation time applied. In this case, it is observed how the model recommends a series of theoretical hours, and the applied time column shows no irrigation activity, indicating that the crop has not been irrigated those days [43]. This type of information can be stored in a structured database since the data can be defined as Integer, Date, Time, Float, Char, Varchar, or String data types, among others.

The most common structured databases are relational databases since they use as a storage base the structure of data groups according to relational algebra. The data query language used is the Structured Query Language (SQL), a standard that allows the definition of data groups and related data types and their manipulation. SQL uses the SELECT statement to perform data searches, INSERT to store information, UPDATE to modify data, and DELETE to delete information. Traditional analytics have focused on structured data as it is possible to access these data through SQL queries [42]. Relational databases include MySQL, PostgreSQL, Oracle, SQL Server, etc.

2.3.2. Semi-Structured Data

Semi-structured data are irregular data that may be insufficient and have a structure that changes as new data are entered; therefore, their structures are unpredictable [44]. This means they are neither table-oriented in a relational database model nor ordered in object-oriented databases. Semi-structured databases enable the storage of data from different sources that have related properties but are different in format. Examples of semi-structured data are email, Extensible Markup Language (XML), JavaScript Object Notation (JSON), Comma-Separated Values (CSV) files, and Key-value, among others [42]. Figure 5 shows semi-structured data in JSON format, as the information is stored as a document and not as a data table. The data are extracted from a satellite image, and then the data are stored as required at that time, for example, coordinate data, the type of analysis, the information about the image, the date, etc.



```
{
  "_id": ObjectId("56649b87a54d75221dd3ac45"),
  "geometry": {
    "type": "Polygon",
    "coordinates": [
      [
        [-124.0,45.0],[-123.5,45.0],[-123.5,44.51],[-124.0,44.5],[-124.0,45.01]
      ]
    ]
  },
  "type": "Feature",
  "properties": {
    "source": "papsin_wfdei.cru_hist_default_firr_set_whe_annual_1979_2012.ne4",
    "centroid": {
      "geometry": {
        "type": "Point",
        "coordinates": [-123.75,44.75]
      }
    },
    "value": {
      "start": [Numberint(1979)]
    },
    "step": NumberInt(1),
    "values": [
      298.6,292.7,311.6,305.6,291.1,346.5,323.4,337.3,328.2,310.8,298.1,316.1,326.0,
      289.5,332.5,302.6,281.7,310.7,272.9,320.7,271.6,370.3,303.9,286.7,295.7,298.2,
      307.7,307.6,309.9,321.0,273.3, 290.7,314.2, null]
  },
  "timestamp": "2015-12-06T14:33:11.030204",
  "simulation": "ast_whe"
}
```

Figure 5. Semi-structured data of a geographic area.

Semi-structured databases enable the storage of the information available at the moment without generating a previous structure, as relational databases do. However, on the other hand, they allow for storing detailed information of an attribute, for example, coordinate values. In this sense, semi-structured databases are scalable and adjustable to the analysis requirements of decision makers. In addition, these databases allow us to work with clustered data volumes and process them quickly. Examples of semi-structured databases are MongoDB, CouchDB, Neo4J, Redis, etc.

2.3.3. Unstructured Data

Unstructured data have no defined structure, as they stores raw files such as bitmaps, images, comments or text, e-mail, and other data types that are not part of a database. Although e-mails are organized in a database, the body of the message contains unstructured text, which must be processed to obtain the information. Examples of unstructured data are files such as Motion Pictures Expert Group 3 (MP3), Portable Document Format (PDF), Joint Photographic Experts Group (JPG), Plain Text Document s(TXT), and Tag Image File Format (TIFF), among others [42]. Table 2 presents some examples of unstructured data.

Figure 6 describes the evolution of unstructured data from traditional numerical data to image data. Numerical data are the simplest since they can be stored in a few bytes of memory (e.g., 2 or 4 bytes). Each character of a text is stored in 1 byte. On the other hand, images and audio require more storage space. In particular, remote sensing images hold much information from sensors. Finally, videos use a series of stored images, requiring more memory space.

Table 2. Examples of sources and unstructured data.

Source	Examples
Radar	Seismic, oceanographic, meteorological, and vehicular.
Static	PDF files, printable files, scanned documents, faxes.
Dynamics	This type is derived from documents that can be created, edited, reviewed, and approved by many people or groups, such as white papers, policy procedures, business documents, and other office documents.
Digital media	Audio, video, animation, and images.
Communication documents	Email, social content, web documents, and instant messaging records
Location/Geo Data	Remote sensing images, sensors, GPS, weather, traffic.
Social Media	Facebook, YouTube, Instagram, LinkedIn.
Sensor	Weblogs, Detail Record, equipment longs, smart meters, manufacturing sensors, trading systems, data records.
Logs	File log, clickstream
Transactions	Web store, Customer information from CRM systems, transactional ERP, general ledger.
Micro-Bloggins	Customer feedback streams, Twitter

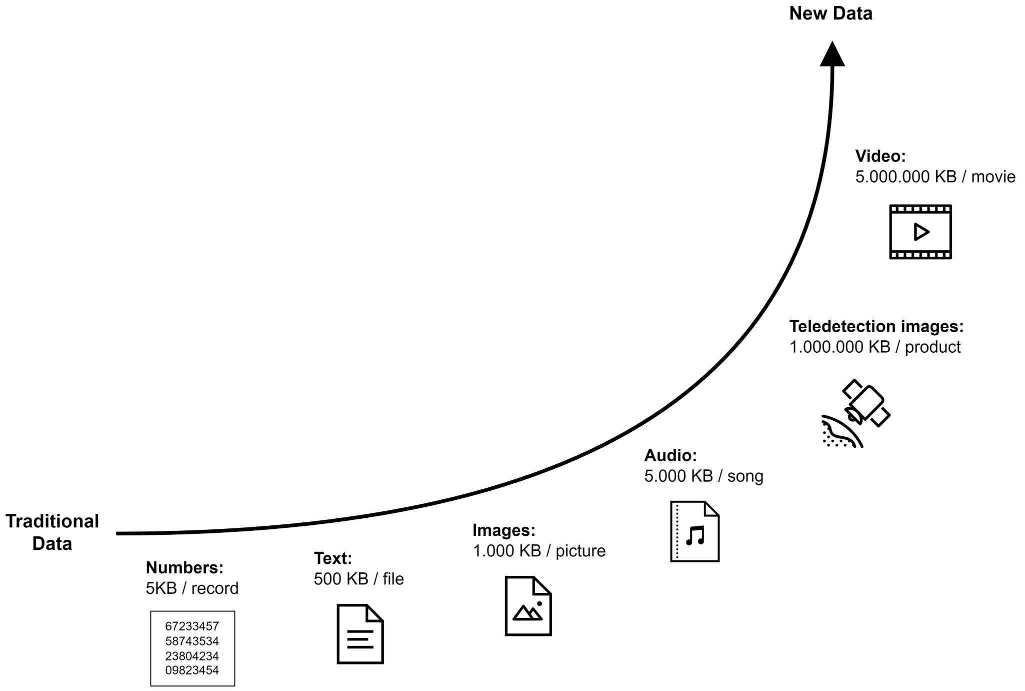


Figure 6. Evolution of unstructured data.

Unlike structured data, unstructured data cannot be easily explored, sorted, or visualized, let alone analyzed. Instead, data transformation tools and procedures are required to extract knowledge relevant to the organization [12]. An example of unstructured data is those coming from satellites circling the Earth, which, through remote sensing, capture data from the planet. These photographs are illustrations of the Earth's character as seen from space. Some examples of data that can be obtained are solar radiation reflected from the Earth, infrared radiation, and backscattered radar intensity. These data are obtained through sensors onboard satellites, which charge radiation of different wavelengths providing multispectral raster data [45]. Figure 7 presents an example.

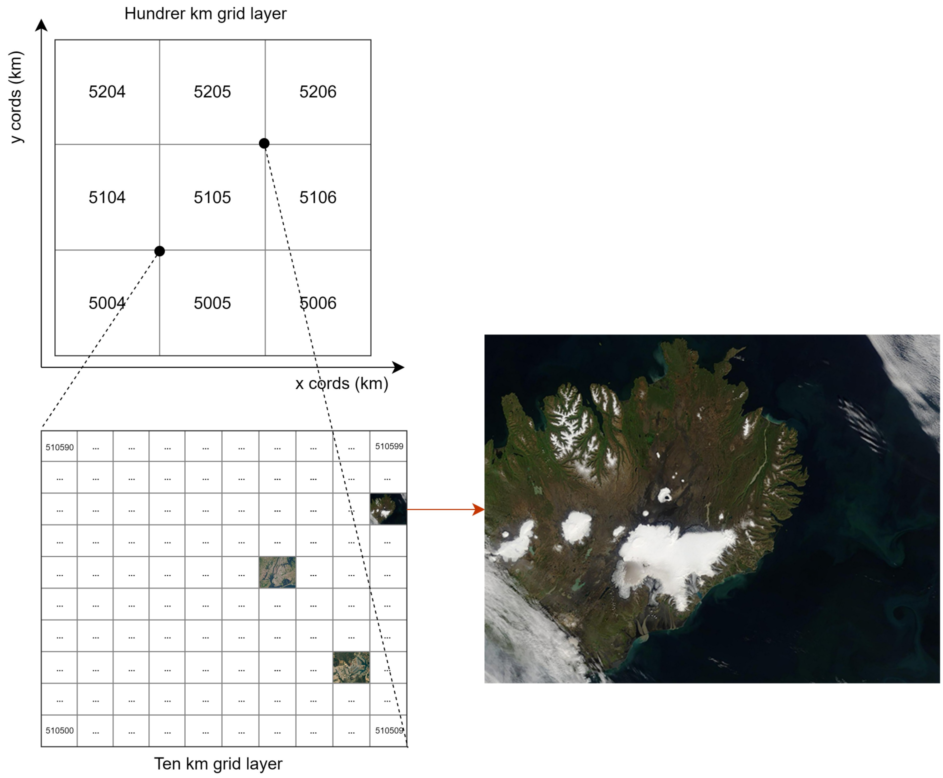


Figure 7. Multi-spectral Raster Data. (Image taken from MODIS).

Multispectral and multidimensional data are typically available in labeled multiband image files (GeoTIFF), an extension of the TIFF structure. For example, a Landsat 7 photograph is a GeoTIFF file consisting of eight spectral bands, individually storing a distinct wavelength spread or ejected from the Earth's exterior. Collecting this data type is a difficult job, as it must be processed quickly through parallel and distributed systems due to the size of the data [45].

Multiband images store multiple values in a single pixel because they store data from multiple sensors concerning the geographic location. Depending on the analysis required, the bands for which information will be retrieved from the sensors are selected. To achieve a more accurate representation of the Earth's surface phenomena, data from different bands can be used [45].

The most commonly utilized active sensors in remote sensing are laser altimeters, radar, and Light Detection And Ranging (LiDAR). Passive sensors operate in the electromagnetic spectrum's visible, infrared, and thermal parts to detect the natural radiation emitted or reflected by objects. Examples of passive sensors are Red, Green, and Blue (RGB), hyperspectral, multispectral, and thermal. In particular, LiDAR sensors obtain 3D properties in the state of point clouds through photogrammetric processing [46]. Information from LiDAR data is used to generate statistics such as mean height, height percentiles, and other data to estimate forest biomass [23]. With the use of ML, it is possible to classify tree species, regardless of the type of sensor used [46].

Satellite imagery and analysis data are currently available from Amazon Web Services [47] and Google Earth Engine (GEE) [48]. GEE uses Hadoop as a platform for data storage, processing, and analytics. These platforms have opened doors to large-scale geospatial analysis and monitoring, as researchers have leveraged these high-performance computing platforms and ML techniques to analyze global changes [49,50].

2.4. Massive Storage

2.4.1. Hadoop

Apache Hadoop is an open-source data processing ecosystem used for distributed computing, which has been created to address Big Data problems. In addition, Hadoop has been extended to use geospatial data. Hadoop generally contains a Hadoop Distributed File System (HDFS) and a MapReduce programming environment for data processing [45]. Figure 8 depicts the architecture of the Hadoop ecosystem [51].

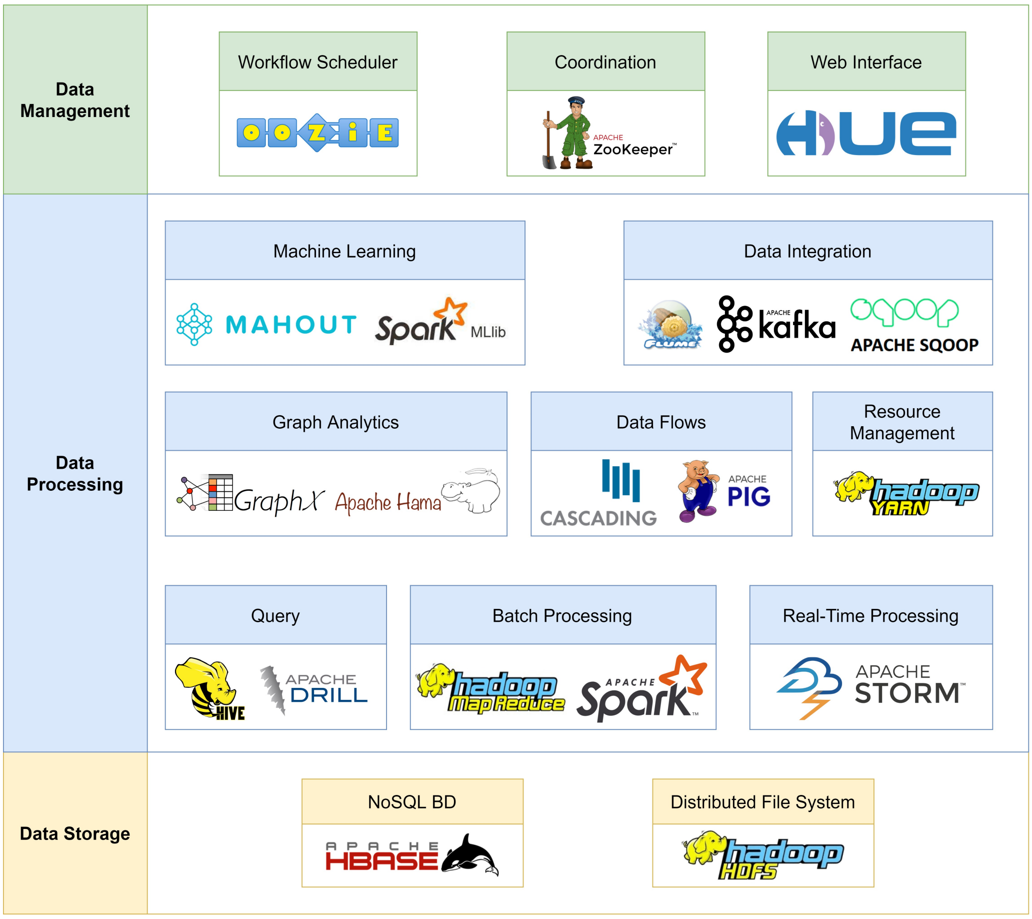


Figure 8. The Hadoop ecosystem.

2.4.2. Cloud Computing

Cloud Computing provides various services over the internet that are scalable. This technology enables the sharing of resources using the infrastructure owned by a cloud service provider. Users or customers of the provider can access resources on demand on a pay-per-use basis. It allows the abstraction of infrastructures, such as storage, network, and applications, through its three services: Platform as a Service (PaaS), Infrastructure as a Service (IaaS), and Software as a Service (SaaS) [52]. The fourth layer of services is Business Intelligence (BI), which contains applications for measuring management indicators. Figure 9 shows all the types of services in Cloud Computing.

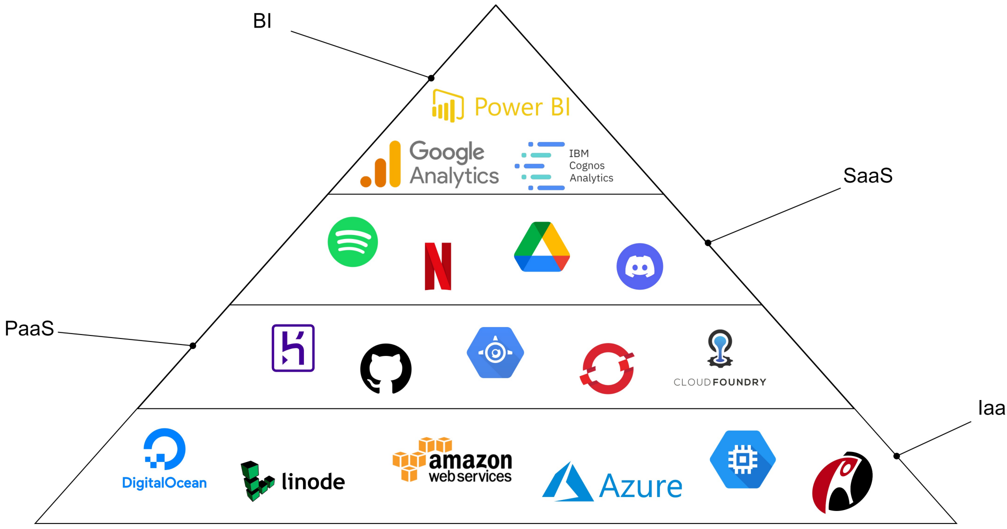


Figure 9. Cloud Computing Services.

3. Methodology

The study method employed was a Systematic Literature Review (SLR). In particular, we use PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) criteria that provide a sequence of steps to follow: identification, screening, eligibility, and inclusion. On the other hand, the methodology proposed by Kitchenham and Charters [19] was used to define the objectives and research questions (RQ). The RQs in this study are the following: (1) What problems are solved with Agricultural Big Data and ML? (2) What data types are used in ML and Agricultural Big Data? (3) What are the sources that generate data in Agricultural Big Data and ML? (4) What are the storage platforms? A search string was constructed in the review process using the keywords selected from the RQs. The terms selected were "Machine Learning," "Big Data," and Agriculture to obtain the maximum possible number of papers to be analyzed. The population-intervention-comparison-results-context criteria were also applied to relate the RQs (PICOC [19]). Agricultural Big Data is the application area and, therefore, the population. The data retrieval process has to do with the intervention. For this study, data matching was not applied. The results are the solutions to the identified problems of implementing ML in Agricultural Big Data.

The work was performed with a team of 4 researchers and an assistant. First, one of the investigators prepared the review protocol plan to ensure the correct selection of studies. Then, two of the investigators carried out the search process. All participated in the filtering process, so it was necessary to hold two meetings weekly to eliminate biases and doubts. Finally, all the research teams read the selected papers to extract the information that provided answers to the RQs.

The search string "Machine Learning" OR "ML," AND "Big Data," AND "agriculture OR farm" was used in Scopus, IEEE, ACM, Springer, MDPI, and Web of Science. The words in the search string had to appear in one of the following sections of each article: title, abstract, and keywords. In addition, conference papers and English scientific journals were considered inclusion criteria. Another criterion was the date of the publications, from 2015 to 2022.

The filtering process was as follows. First, 5620 articles were identified in ACM, 536 in the Web of Science database, 1321 in Springer, 56 in MDPI, and 122 in IEEE. The research team applied the exclusion criteria to eliminate papers unrelated to this study. First, duplicate papers, or papers not written in English, are eliminated. Then, all researchers read the title and keywords to select papers containing any of the search words. Next, all researchers read the abstracts to determine the contributions in the field of Agricultural Big Data. Finally, each article was read to eliminate those not using ML. In the end, 43 documents were obtained that met the expectations.

A total of 7655 articles were reviewed. First, screening eliminated duplicates, books, and technical papers, resulting in 5233 potential articles. The next step was to read the abstracts of the articles to check if they contained information on Agricultural Big Data. After this process, 521 articles remained. These papers were reviewed in their entirety to check if they included information on Big Data and ML in agriculture. Many of the articles are reviews or editor's comments. In the end, 43 papers were relevant. Figure 10 summarizes the process followed by the team to select relevant papers, and Figure 11 presents the selected papers by year.

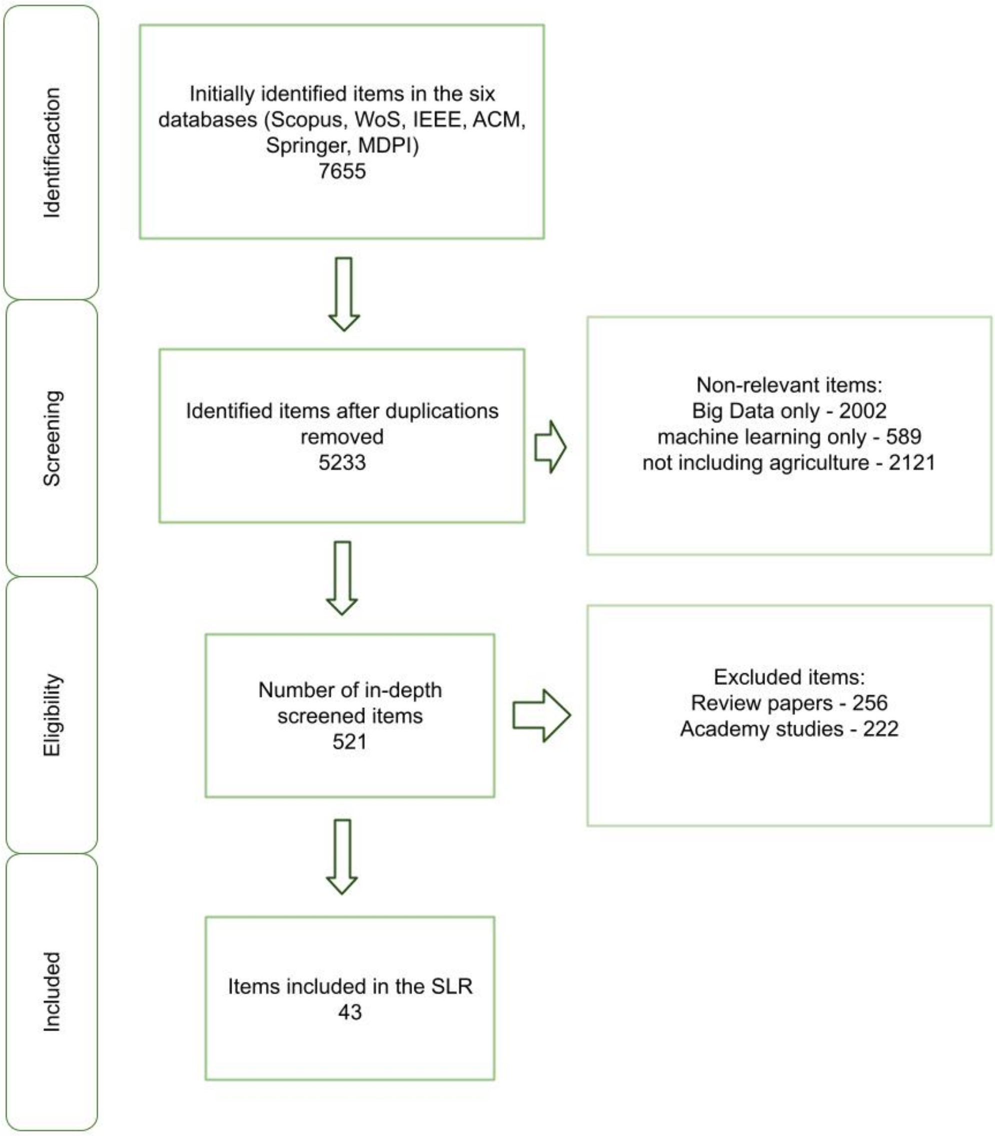

Identificaction
Initially identified items in the six databases (Scopus, WoS, IEEE, ACM, Springer, MDPI)
7655
Screening
Identified items after duplications removed
5233
Non-relevant items:
Big Data only - 2002
machine learning only - 589
not including agriculture - 2121
Eligibility
Number of in-depth screened items
521
Excluded items:
Review papers - 256
Academy studies - 222
Included
Items included in the SLR
43

Figure 10. Selection process for relevant papers.

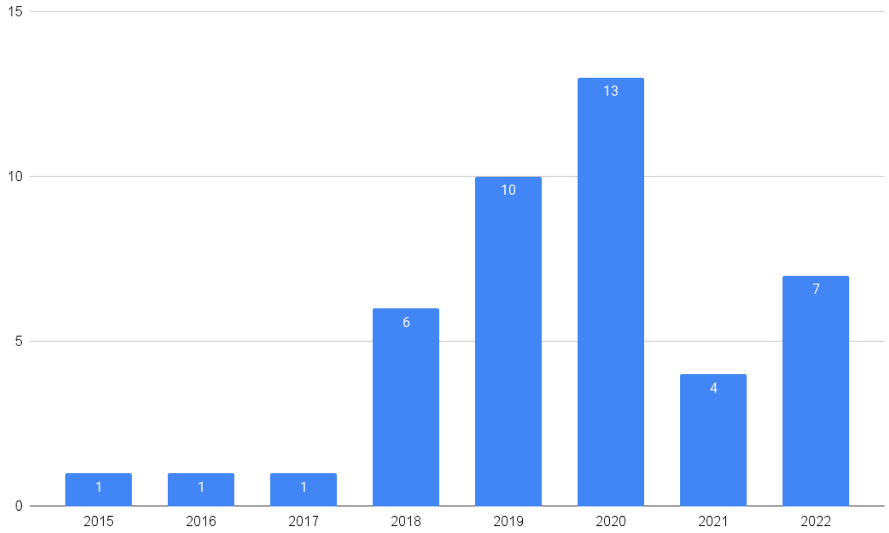


Figure 11. Selected papers by year.

4. Results

This section answers the RQs described in the methodology.

4.1. *Agricultural Big Data and Machine Learning*

The 43 selected articles describe various problems in the domain of agriculture that are possible to solve through Big Data and ML. Many Agricultural Big Data solutions enable crop improvement, animal production management, weather, land management, food availability and security, climate change management, and weed prevention, which are essential for farmers' decision making. Figure 12 shows the number of papers found by problem and area.

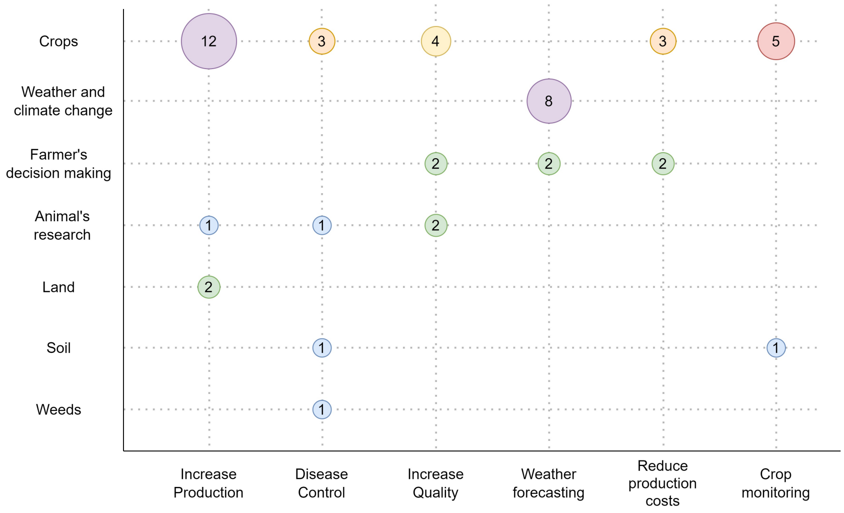


Figure 12. Problems vs. areas in Agricultural Big Data and Machine Learning.

The Figure shows that most of the selected papers solve problems related crops, weather and climate change, and farmers' decision making. The main concern in crops is improving production, reducing costs, increasing quality, and managing infections. Balducci et al. [53] propose using ML to analyze environmental data such as temperature, wind, and humidity, as well as productive and structural data such as land area and soil type. In the same vein, Kedarmal et al. use a graph for farmers to take a convenient step to enhance production by predicting crop yields. To do so, they rely on an ontology of smart agriculture through ideas and effects related to agriculture, which are linked in a knowledge graph [54]. In contrast, Ramaraj et al. [55] analyzed the various rice crops for yield prediction. The data used are crop morphological characteristics and yield parameters. Yoki et al. developed the BMS (Big Data Application Machine Learning-based Smart Farm System) to improve farmers' income by increasing crop productivity [56]. Priya et al. [57] present a farming representative to maximize crop production and quality by suggesting types of crops to be planted according to soil conditions.

In the case of crop identification and classification, Shelestov et al. [58] used satellite images to generate a crop classification map to analyze each zone separately. The same problem is addressed by Yahata et al. [59], who use detection techniques on photographs acquired from a cyber-physical system to obtain crop growth data and environmental information. The authors developed good rules to define a good harvest by detecting flowers and pods. Saldana and Guo [60] automatically classified four types of fruit trees (coconut, banana, mango, and papaya) using high-resolution aerial images. They used ML to locate and classify trees and locate paths for transport. They used algorithms to decide the density of clusters of individual tree species delivering a more helpful understanding of the condition of the agricultural site. Finally, they developed a method to optimize fruit picking founded on distinctive scenarios, such as maximum time, path length, storage location, and safety points.

Fenu and Malloci [61] use ML models to improve the detection of crop diseases, mainly late blight disease in potatoes. For this, they used data from weather stations in the region. Finally, Tombe [62] uses images of the crop to determine the state of health and identify diseases and weeds. It integrates computer vision for smart farming.

Regarding farmers' decision making, Dutta et al. [63] captured data from domain experts on agricultural processes, soil understanding, and harvest optimization about climatic conditions. Then, they analyzed the data to develop low-cost plantations.

Doshi et al. [64] developed AgroConsultant, a Big Data system that helps farmers make better decisions on which crops to grow, taking into account the geographical location of their operation, the planting season, soil characteristics, and environmental factors such as temperature and rainfall.

Rehman et al. [65] used sensors to capture real-time data. These climatic data from the soil and atmosphere allow planning planting dates. The third problem concerns yield enhancement. The work of Tarik and Mohammed [66], who predict the cereal production rate in an unstable climate area, stands out. They captured water data through a Big Data system. The data are processed to obtain dew point, humidity, temperature, wind direction, wind speed, pressure, visibility, gust speed, events, precipitation, weather conditions, etc.

In Kumari et al. [67], the authors proposed a multidisciplinary model for smart agriculture using the IoT, Big Data, Cloud Computing, Machine Learning, Wireless Sensing Networks, Sensors, and Mobile Computing. Data will be collected through sensors, including data on rainfall, temperature, humidity, soil, and moisture. These data are analyzed to send management indicators to farmers and agricultural product vendors. The arrows relate to weather conditions and temperature. The authors conclude that the real-time system will increase productivity and reduce costs.

Katyayan et al. [68] used sensors in the field to collect personalized data in real-time. The data are managed through analytics to provide the user with accurate, personalized assistance and guidance solutions. For the analysis, they used Natural Language Processing and computer vision. Additionally, they collected data from real-time crop images analyzed with ML. Finally, they use an embedded system that accumulates historical information to be analyzed by experts, helping farmers with decision making.

Melgar-García et al. [69] present a new big data triclustering approach based on evolutionary algorithms to extract three-dimensional patterns. They use the vegetation index of grapevine crops to carry out precision agriculture. The aim is to develop specific crops considering each growing area's variability.

Wang and Mu [70] propose a modern agricultural technology platform. The platform is for precision agriculture through a network of wireless sensors, which help measure wavelengths of light to determine wheat yellow rust. Additionally, they use aerial photographs from drones to analyze the disease's severity and the outbreak's area.

Wang et al. [71] developed an intelligent agricultural system to process the massive data of agricultural monitoring. The system is divided into three sub-systems: the remote sensing agricultural situation monitoring and information service system, which includes growth monitoring and yield estimation; the remote sensing system for monitoring the resource environment and natural disasters; and the monitoring and supervision system for new rural development and key agricultural construction, which includes new rural construction and low-yield fields. Ahamed [72] uses remote sensing for diversified land use planning for food and nutrition security, crop growth monitoring, yield forecasting, land suitability analysis, forest productivity, and drought assessment for crops, vegetables, and fruits.

Venkatesan et al. [73] propose an ML-based prediction model to analyze the energy use of environmental devices to collect data on crop growth in a smart paprika farm through the IoT. The authors propose a multi-time-randomized energy management plan for greenhouses with energy sources. They also use ARIMAs, which are integrated models of climate and crop production.

Kuchler et al. [74] developed a classification strategy for the annual mapping of integrated crop–livestock systems (iCL) on a regional scale. The authors discuss the efficient use of land, indicating that it is important to use the land for crops and livestock. They show that this is possible through intelligent analysis.

4.2. Data Types

The data used in the different Agricultural Big Data proposals come from sensors, which, in some cases, are IoT devices, satellites, cameras, GPS, databases, and data from farmers' expert knowledge. Figure 13 shows the primary data used using a cloud of concepts. It is observed that the most used data in the selected papers are temperature, humidity, crop area, wind direction, and wind speed.

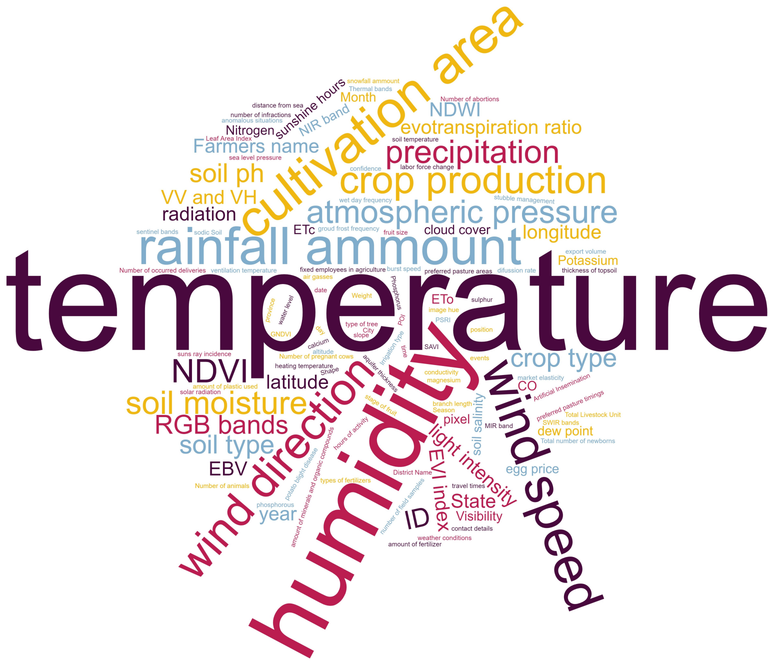


Figure 13. Cloud of data concepts used in Agricultural Big Data.

The data mainly used are structured data. Several papers mention using unstructured data from photographs, GPS, and satellite images. Table A1 in Appendix A presents a summary of the data by the author.

The following subsections describe the use of structured and unstructured data.

4.2.1. Structured Data

Structured data have a defined format, such as Int, Date, and Float. In addition, 93% of the selected papers describe the data and their use. Tables 3–7 present a summary of these data. The following subsections describe the use of the data according to the data source.

Table 3 presents the structured data whose sources are sensors.

Table 3. Structured data from sensors.

Author	Data	Type of Data
[53]	Temperature average, min and max, rainfall amount, amount of phosphate and potash minerals, humidity, geo-coordinates (station id, point of presence (Poi), latitude, longitude, altitude), sun's rays incidence (r_inc), wind speed and direction, atmospheric pressure, etc.	Int
[57]	Air temperature, relative humidity, wind speed, wind direction, soil temperature, soil moisture, radiation, diffusion rate, and precipitation.	Int
[75]	Hours of activity, travel times, preferred pasture areas and timings, anomalous situations, number of fence and posture infractions.	Int, String
[76]	Plant data: temperature, humidity, illumination intensity, air gasses (CO_2, O_2, O_3, NO_2), plant ID, time; system user data: such as name, password, and role; video file data: device, location, time, access port.	Int, DateTime, String
[77]	Air temperature, air humidity, light intensity, soil moisture, fruit size, branch length, soil salinity, wind speed, wind direction, conductivity, soil temperature, CO_2, PH.	Int, Float
[34]	Temperature, rainfall, PH, water level, nitrogen, phosphorous, potassium, calcium, magnesium, sulfur	Int, Float
[54]	Nitrogen, phosphorus, potassium, saline soil (ha), sodic soil (ha).	Int, Float
[65]	Temperature, humidity	Int, Float
[66]	Temperature, dew point, humidity, pressure, visibility, wind direction, wind speed, burst speed, events, rain, weather conditions, month, day.	Int
[56]	Color, weight, inner temperature, outer temperature, humidity.	String, Int, DateTime
[67]	Temperature, humidity, moisture, and rain.	Int, Float
[78]	Temperature, relative humidity, mean sea level pressure (hPa), snowfall amount, sunshine duration, evapotranspiration, FAO reference evapotranspiration, wind speed, wind direction, soil temperature, soil moisture.	Int, Float
[73]	Internal temperature, internal humidity, ventilation temperature, heating temperature, outside temperature, outside solar temperature, dew point, hourly accumulated light, hourly solar radiation, temperature difference, crop output production	Int, Float, String

Bendre et al. [79] describe the technologies used in Precision Agriculture and the data types when Big Data is required. He explains that the data sources are usually agriculture machinery data, remote sensing data, GIS data, GPS data, and farmer documents, such as crop monitoring and Varying Rate Fertilizers. The authors used daily data of minimum and maximum temperature (in oC), humidity (in %), and precipitation (in mm) from a weather station to predict the climate. All these structured data are of type Int.

On the other hand, Balducci et al. [53] use three different data sets from enterprise, scientific investigation, and national statistical academies. With the IoT and ML, they obtain data for the forecasting and reconstruction of cutting or erroneous data, as agreeably as the detection of defective hardware sensors from the monitoring stations. The data from the 41 monitoring stations are a mix, max, temperature average, precipitation amount, a portion of phosphate and potash minerals, humidity, and wind.

Priya et al. [57] use data from satellite imagery, sensors located in fields, irrigation-related messages, and climate and crop data. Some attributes are radiation, relative humidity, wind speed, air temperature, soil temperature, wind direction, soil moisture, diffusion rate, and precipitation. All these structured data are Int. The data are used to analyze each area's most suitable crop type.

Nóbrega et al. [75] use a system of rules to monitor sheep in vineyards. Structured data are obtained by processing sensor data through ML techniques and validation rules. The structured data are hours of movement, travel times, select range areas and timings, unknown problems, and several fence and posture infractions. The select pasture areas and anomalous conditions data are STRING data types or numerical codes assigned according to the rules. All other data are of type Int.

On the other hand, Yang et al. [80] use a data sensing instrument deployed on a farm to organize plant environmental and growing data in real-time by using different sensors to accumulate data such as humidity, temperature, air indices, and illumination. The data are then uploaded to a cloud forum. The purpose is to evaluate the growth efficiency of plants at every moment to direct agricultural activities in smart greenhouses. All structured data are of type Int, except the time, which should be of type DateTime. The system also stores user data, such as name, password, and role, and video data, such as device, location, and time. Most of these data are of the String type. Although the data are stored in MySQL, when a data volume is generated, the data are transferred to the platform with Hadoop. A similar Agricultural Big Data system was developed by Wang et al. [77], in which they used sensors to analyze the growth of pears. The data collected are air temperature, humidity, light intensity, soil moisture, fruit size, branch length, and soil salinity. Wind speed, direction, conductivity, soil temperature, CO_2, and PH, are Int or Float structured data. On the other hand, Alex et al. [34] use sensor data to collect information on fertilizers and how they influence plant growth. The data used are structured data of type Int and Float, despite storing them in Hadoop.

Tarik et al. [66] used data from sensors located at a weather station. The data are temperature, dew issue, pressure, humidity, visibility, wind speed, wind direction, gust speed, events, weather conditions, precipitation, etc. All these data are Int. In Donzia and Kim [56], they use a smart farm system to monitor the crop environment, accelerate climate resilience, and increase production through real-time knowledge of agricultural regulations. They use sensors to capture data such as color, weight, indoor temperature, outdoor temperature, and humidity. Color is a String data type stored as Int when performing the digital conversion. All other data are Int. Table 4 presents a summary of data and data types obtained from cameras.

Table 4. Structured data from photographic cameras.

Author	Data	Type of Data
[59]	Climate, temperature, humidity, solar radiation, soil condition, shade.	Int, Float, String
[81]	pixels	Int
[68]	Temperature, light, humidity, and soil moisture.	Int, Float

Yahata et al. [59] use two methods to observe soybean flowers and pods in natural fields. The developed image detection methods include sensors in an agricultural cyber–physical system to analyze plant growth status and environmental knowledge (e.g., weather, humidity, temperature, soil condition, solar radiation, etc.). The data obtained are analyzed using rules useful for proper cultivation. On the other hand, the images are processed and analyzed using ML. Flower detection uses a coarse-to-fine method, where flower prospect locations are first detected using information from tonality and Simple Linear Iterative Clustering. Data used are Int, Float, and String Type. On the other hand, Vasumathi and Kamarasan obtained a test data set from images extracted from a camera and a cell phone. They captured more than 200 images of different disease-free fruits. They developed patterns to identify three levels of diseased fruits, low, medium, and high, all through pixel comparisons [81]. The pixel data of the images are Int. However, the fruit condition data were stored in a NoSQL database.

As for the structured data from GPS and databases, Tables 5 and 6 present a summary.

Table 5. Structured data coming from GPS.

Author	Data	Type of Data
[82]	cropland class, number of field samples	Int

Amani et al. [82] used a vehicle's GPS to determine the crop type. It allowed them to easily and quickly obtain numerous field samples of various types of crops.

Table 6. Structured data from databases.

Author	Data	Type of Data
[57]	Historical crop data.	String, Int
[83]	max temp, min temp, precipitation, month, year	Int,Float, Date
[61]	temperature, relative humidity, wind speed, wind direction, precipitation accumulation, solar radiation, potato blight disease, cultivar resistance	Int, Float, String
[55]	crop yield, farmer's name, Place–Village–Panchayat–Taluk–District, survey number, contact details, soil type, paddy variety, types of fertilizers used at the different stages of cultivation, field size, irrigation type	String, Int, Float
[64]	soil type, aquifer thickness, soil pH, thickness of topsoil, precipitation, temperature, latitude, longitude, distance from sea, rainfall, production area, under, cultivation	Int, String
[84]	soil pH, state of Shire, winter crop, amount of cultivation, stubble management	Int, Float
[60]	type of tree, x (pixel), y (pixel), size, confidence, land, cover, class, ground cover: impervious surfaces, buildings, low vegetation, trees, cars, and background	Int, String
[85]	max air temperature, min air temperature, average temperature, relative humidity, wind speed, solar radiation, sunshine hours, reference evotranspiration	Int, Float
[86]	min temperature, max temperature, avg temperature, humidity, wind speed, precipitation, wind direction, cloud cover, visibility, atmosphere pressure	Int, Float
[87]	precipitation, temperature (minimum, maximum and average), cloudiness, vapor pressure, frequency of wet days, frequency of frost on the ground.	Int, Float
[88]	consistency of all animals, percentage of artificial insemination, mean value of age of cows expressed in days, mean number of parturitions per cow, number of occurred deliveries, mean number of necessary inseminations which resulted in positive pregnancy diagnosis, etc.	Int, Float
[82]	NDVI, NDWI	Int
[89]	wgg price, duck egg price, export volume, output, market elasticity, labor force change and inventory as variables.	String, Int, Float
[90]	crop field dataset, name area	String, Int, Float, Date

Balducci et al. [53] describe data from the Istat database, which contains well-structured, publicly available data used to predict future crop quantities. In addition, scientific data from the National Research Council predict specific crop species. The data used are soil type, year of the time series, crop type, province, total altitude location, cultivation area, full crop production, and total harvest production. Soil type, crop type, and province are String-type data; the date is of type Date, and the other data are of type Int. One of the challenges cited by the authors is the reading of data from the National Research Council since many of these are incomplete and only partially sorted.

In Sathiaraj et al. [83], they processed data from the Applied Climate Information System (ACIS) to obtain the climatic elements such as precipitation and minimum and maximum temperature.

Fenu et al. developed DSS Land, a system that acquires meteorological data from ARPAS (Regional Agency for Environmental Protection of Sardinia) weather stations according to specific procedures and time intervals. The stations observe several meteorological variables, including relative humidity (%), the direction of the accumulation of precipitation (mm), temperature (oC), wind speed (km/h), and solar radiation (W/m^2) [61]. All these data are of types Int and Float. They also process the data to classify the cultivar resistance, a String data type.

Gnanasanka and Ramaraj [55] use a soil data set, a weather data set, a rainfall data set, and a pest data set. These datasets are extracted over the entire rice growing period (from seed sowing to harvest time). Initially, the meteorological data sets of rainfall and weather were extracted from CTC-CECRI, Mandapam, and the State Groundwater Resources Data Center in Chennai. Subsequently, they conducted soil and water tests to know the soil types and their moisture and evaporation level and water tests to identify the pH value and type of water (hard water or soft water). All these data are structured from relational databases that store the data from IoT devices.

In Doshi et al. [64], they use data from the "India Agriculture and Climate Data Set" database to generate training datasets for ML. The analyzed data allow generating alternative recommendations for farmers. The database contains soil and meteorological data from 1957 to 1987. The parameters refer to five major crops (jowar, bajra, wheat, maize, and rice) and 15 minor crops (jute, cotton, barley, gram, groundnut, tobacco, potato, ragi, tur, sesame, rapeseed and mustard, soybean, sunflower, sugarcane, and pulses). The data analyzed are soil type, aquifer thickness, pH, topsoil thickness, rainfall, temperature, and location parameters. All these data are Int and Float. The second database used is "Latest Socio-Economic Statistical information & Facts About India," from which they obtain rainfall data for 117 years. With these data, the authors could predict the sector's rainfall [64].

Ip et al. [84] use data from two databases to analyze weeds. The first database contains information from herbicide resistance tests and annual ryegrass samples from farms in southern Australia. The second database comes from farm surveys of winter crops and the amount of cultivation before sowing. It also contains information on stubble management, the amount of water since sowing, and soil pH. All these data are of type Int and Float.

On the other hand, Saggi et al. [85] used data from the India Meteorological Department database to calculate evapotranspiration accurately, as it recreates a vital part of irrigation water scheduling for its efficient use. The data are maximum atmosphere temperature, relative humidity, min air, average temperature, wind speed, solar radiation, reference, evapotranspiration, and sunshine hours,. All data are of type Int or Float. Reddy et al. [86] also use Indian Meteorological Department data to predict rainfall behavior. The data used are min temperature, max temperature, wind direction, avg temperature, humidity, precipitation, cloud cover, wind speed, visibility, and atmospheric pressure, which are Int and Float data types.

Abbona et al. [88] use genetic programming to analyze the health and growth status of the animals. To do this, they access data from the ANABORAPI system, which contains the history of all the farms registered in the Herd Book of the Breed. ANABORAPI stores information on each farm, which technicians record during routine checks by veterinarians and, presently, by agriculturalists. The data are recorded through a Workabout portable computer or with a smartphone. The data obtained are the average value of cow age expressed in days, the consistency of all animals, the percentage of artificial insemination, the average numeral of calvings per cow, and the total number of calvings that have occurred. In addition, the averages of the required inseminations that resulted in a positive gestation diagnosis have been obtained. Several of these structured data are of type Int or Float.

Amani et al. [82] use the database of the federal Department of Agriculture and Agri-Food Canada, which provides maps of the Annual Inventory of Crops in Space. The maps allow identification of the non-agricultural land cover within the agricultural area of Canada. The authors report that spatial remote sensing is one of the forms to obtain direct reports clarifying the evolution of the atmosphere due to the breadth and complexity of agricultural regions.

Su and Wang [89] used egg data to analyze the factors influencing prices. The data include egg price, duck egg price, export volume, output, market elasticity, labor force change, and inventory. The data are of type String, Int, and Float, as they come from a structured database.

Table 7 below presents a summary of structured data from satellites. The data obtained have been processed through various systems, mainly GIS and GEE.

Table 7. Structured data from satellites.

Author	Data	Type of Data
[87]	precipitation, min temperature, max temperature, avg temperature, cloud cover, vapor pressure, wet, day frequency, ground frost frequency, color bands	Int, Float
[82]	NDVI , NDWI.	Int, Float
[91]	NDVI, NDWI, PSRI, B8A, B12, B11, B08, B07, B06, B05, B04, B03, B02	Int, Float
[92]	RGB bands, NIR band, SWIR 1 and SWIR 2 bands, thermal bands, NDVI index, EVI index, NDWI index, latitude, longitude, slope	Int, Float
[69]	three-dimensional datasets, NDVI index, Soil-Adjusted Vegetation Index (SAVI), EVI Index, and Green NDVI (GNDVI)	Int, Float
[74]	NDVI, EVI, near-Infrared spectral band (NIR), and mid-Infrared spectral band (MIR).	Int, Float

It is possible to obtain satellite images containing valuable information for the crop, climate, and land analyses. Some examples of data are the Normalized Difference Vegetation Index (NDVI), which is possible to calculate using Equation (1); the Normalized Difference Water Index (NDW) and the Normalized Difference Water Index (NDWI), which are given in Equation (2); and the Normalized Difference Vegetation Index (PSRI), which is given in Equation (3):

$$\text{NDVI} = \frac{nir - red}{nir + red} \tag{1}$$

$$\text{NDWI} = \frac{nir - swir}{nir + swir} \tag{2}$$

$$\text{PSRI} = \frac{red - blue}{red\_edge} \tag{3}$$

On the other hand, the Enhanced Vegetation Index (EVI) is used to identify the density of vegetation, including forests. EVI is an optimized vegetation index that detects sensitive vegetation not seen by the NDVI. In addition, the NDWI was used to identify cropland [92]. An enhanced version of the NDVI in Green NDVI (GNDVI) provides information on the land cover status and trends in land cover change, such as degradations reflecting impacts on forage quantity and quality [93]. Other vegetation indices used are: Soil-Adjusted Vegetation Index (SAVI), Transformed SAVI (TSAVI), Modified SAVI (MSAVI), Modified Transformed SAVI (MTSAVI), Optimized SAVI (OSAVI), and Generalized SAVI (GESAVI) [94].

Other data that are possible to obtain through sensors are wind speed, wind direction, soil temperature, radiation, diffusion rate [57], precipitation, min temperature, max temperature, cloud cover, vapor pressure, wet, day frequency, ground frost frequency, and color bands [87].

4.2.2. Semi-Structured Data

Few papers describe the use of semi-structured data. Some authors use them to capture information from sensors that must be processed immediately to obtain the appropriate values. In Nóbrega et al. [75], they use an IoT sensor network to analyze the movement of sheep. The system uses a set of rules that indicate the activity performed by the sheep. What is of interest is that the sheep only eat grassland and not the vineyards. The sensors transmit information to a central system to analyze the preferred pasture areas and timings, anomalous situations, number of fences, and posture infractions. All these data are semi-structured because they must be stored in a documentary database.

The sensor data are processed to obtain the rules. They then use a gateway that incorporates a module capable of mapping the collected information into JavaScript Object Notation (JSON) data structures. This allows the non-IP network and the IP-based Internet to be integrated for rapid analysis by higher-layer applications [75]. The system combines the process that handles real-time traffic and non-periodic traffic. The processed data, for both cases, are stored in a PostgreSQL relational database. Table 8 presents a summary of these data.

Table 8. Semi-structured data.

Author	Data	Type of Data
[75]	Anomalous situations, preferred pasture areas and timings, posture infractions, and number of fences	JSON
[83]	Daily climate datasets from the Applied Climate Information System (ACIS)	Clave-valor
[54]	Crop year, season, state name, district name, crop, area, seasonal rainfall, production	RDF triple store
[81]	Stage of deseased fruit	JSON
[95]	State, city, name, state ID, owner's name	Spreadsheet

Sathiaraj et al. [83], from ACIS, collected climate data. The data were grouped by months or years to develop three sets of climatic data: annual, monthly, and TEF. These were stored in REDIS, a NoSQL database, allowing for in-memory data management. Therefore, the system provides quick access.

Choudhary et al. [54] collect data from India's Open Government Data Platform (OGD) and Rajasthan. The data used are the name of the state, the name of the district, the growing season and year, the area, the production, the seasonal rainfall, the phosphorus, the potassium, the nitrogen, the saline soil (ha), and the sodic soil (ha). The data are stored in a graph database in RDF triple store format.

In Aiken et al. [95], they stored data from Brazilian sources on (beef) cattle from two massive datasets: 44,566 farms purchased from an animal food business and 32,776 processed cattle at a meat packing company. The raw data had to be processed, cleaned, and formatted for storage in a structured database. The data they used to analyze and classify the cattle were state, city, name, state ID, and owner's name.

4.2.3. Unstructured Data

As shown in Table 9, it is possible to obtain data from Sentinel-1A through an active Synthetic Aperture Radar (SAR) sensor, providing C-band images in both single polarization VV and dual polarization VH, which is not affected by cloud cover or lack of illumination. Moreover, the images are freely accessible to any user [58,82].

In Nóbrega et al. [75], they use video data to learn about sheep behavior in a vineyard. To do so, they analyze each video to obtain a set of rules processed by experts. Yang et al. [80] also use video data to monitor the growth status of plants. The videos are stored locally on the farm for only seven days, then stored in a Cloud system. Users can decide to watch the documented or real-time video; the cloud medium obtains the related video submissions to show the association with the connected video device and send the video data to the user. Rehman et al. [65] use various types of data for crop analysis. For example, the real-time data stream comprises soil temperature, atmospheric temperature, and humidity, and structured data from sensors. On the other hand, the practice plan is formulated with information extracted from heterogeneous data sources: (1) real-time data streams from sensors; (2) sensor location information; (3) extraction of information from five-year-old data; (4) government climate forecast; and (5) irrigation system.

Table 9. Unstructured Data.

Author	Data	Type of Data
[75]	Referred pasture areas and timings, anomalous situations, patterns of movement, food preferences.	Videos
[80]	Videos of the state of growth of plants.	Videos
[65]	Text, Web data, CSV	Spreadsheets, Web Data, CSV
[63]	Satellite images, domain knowledge	Spectral images, RDF, URI,
[59]	Imágenes	RGB image
[60]	Drone images, digital surface models, high-resolution UAV imagery	RGB image
[87]	Satellite images, archives	RGB image, CSV, GeoTiff
[82]	SAR images SGX dual-polarization RADARSAT-2 in the mode Wide, optical images Landsat-8, VV, VH	Spectral image, RGB, ASCII, HDF
[58]	SAR images, VV, VH polarizations	Spectral images, ASCII, HDF, GeoTiff
[91]	Optical images, SAR images, VV and VH coefficients	Spectral images, RGB
[92]	Satellite images	Spectral images, ASCII, HDF, GeoTiff
[67]	CSV	CSV
[68]	Images, video, audio, telemetry data, user data	Documents, RGB image, CSV, videos
[69]	Images	RGB image
[71]	Images, GPS, documents data	RGB image, points, PDF

In Dutta et al. [63], they use heterogeneous data from different sources to generate a knowledge system of agricultural processes in conjunction with environmental processes. The data are obtained from a network of sensors, satellite images, large-scale simulated models, meteorological data, and domain knowledge and experience to improve decision making. With the data obtained, the authors developed a Big Data system that incorporates unstructured, undocumented, and ad hoc knowledge into a structured rule base that is directly used to improve the decision support system. In addition to farmers' knowledge data, they use other unstructured data from NASA MODIS, NASA LANDS, and Australian Digital Elevation.

On the other hand, Sumalatha and Akila [87] use historical data from the government portal "indiawaterportal.org" and satellite images for histogram analysis to identify the color pattern on the ground. They analyze data from various environmental features that have affected the terrain. The data used come from CSV files and RGB images.

Amani et al. [82] also use satellite imagery to obtain information on terrain characteristics. The data are extracted directly from the Google Earth Engine (GEE) cloud computing platform, as it allows for improved data processing efficiency from a time and cost perspective. The authors explain that accuracy levels are slightly lower than when using data from the Department of Canada's database. However, GEE will improve this aspect in the future. GEE contains freely accessible remote sensing datasets and various classification algorithms, which can be accessed for different applications on arable land. In addition, Gumma et al. [92] used Landsat data and ML algorithms in the GEE software platform. The authors created a mega-file of 31 bands for five agro-ecological zones in South Asia, which formed a database for image classification and analysis.

GEE stores raster data sets from NASA, as well as a complete Landsat archive, in addition to data that come from the European Space Agency. They use JavaScript or Python to analyze multi-temporal data on a continental scale, which can be shared with other researchers. This mitigates the barriers to using supercomputers in geospatial analysis [92].

Yahata et al. [59] and Saldana and Guo [60] use photographic images to analyze the condition of vegetation, mainly trees. The raw data are RGB images that need to be processed to obtain color and pattern information.

Sitokonstantinou et al. [91] use data from the Umbrella Sentinel Access Point, developed by the National Observatory of Athens, to map rice crops. The authors complement the data with satellite images and photos. They are preprocessing Sentinel-1 data, including (i) cropping of the area of interest, (ii) radiometric calibration, (iii) speckle filtering using the Lee filter, (iv) terrain correction using the Shuttle Radar Topography Mission (SRTM) 10 m, and (v) conversion of the backscatter coefficient ($\sigma 0$) into decibels (dB). The Sentinel-2 images consisted of 13 spectral bands with 10, 20, and 60 m spatial resolutions. In addition, the authors calculated three vegetation indices to improve the feature space, NDVI, NDW, NDWI, and PSRI.

4.3. Data Generation for Agricultural Big Data

Figure 14 shows the number of uses of different data sources for the generation or collection of Agricultural Big Data. The identified sources were categorized into six groups. These are Sensors, Cameras, Databases, GPS, Satellite, and People. Each of the groups is described below.

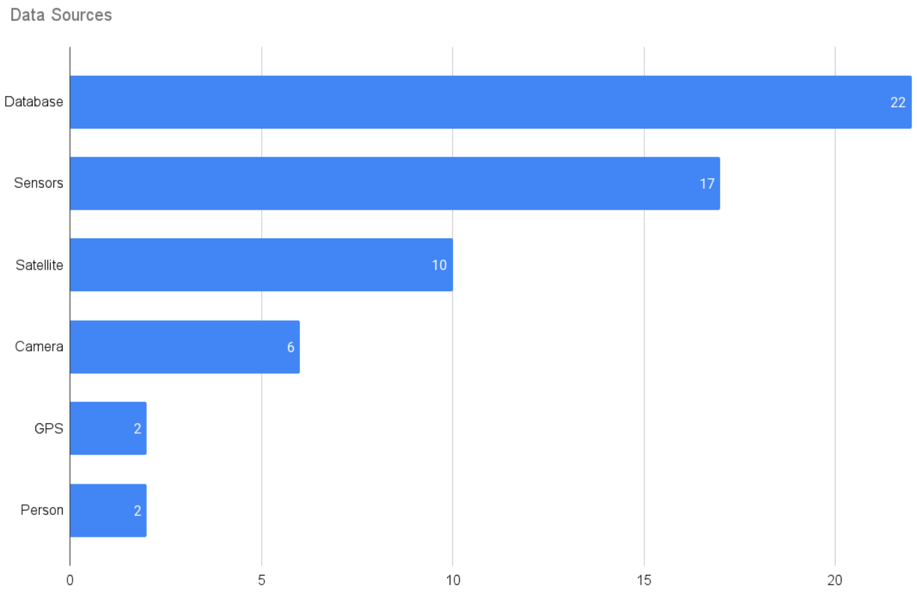


Figure 14. Data sources used in Agricultural Big Data.

The People group includes all sources whose data comes directly from people and their domain knowledge. It is one of the least used sources in the articles analyzed, used in only two papers. Capturing data from this source is a complex task, as the data are neither structured nor documented and exist only in experts' minds. This nature of the data makes them inaccessible to a computer system in a straightforward manner, differentiating them from the rest of the proposed groups. In [63], the complexity of this data source is addressed, and a protocol is offered to systematically capture domain knowledge and integrate the data obtained with other data sources to generate a decision support system.

A Global Positioning System (GPS) is a system composed of satellites and receiving devices in such a way as to identify the receiver's position on the Earth with an accuracy of up to centimeters. This data source, like those in the People group, was only used in two articles. These systems only allow the collection of geospatial data such as the latitude and longitude of the receiving device. In the analyzed papers, the data are collected through devices, including GPS receivers, used during field campaigns to determine the position of certain areas or sections of a terrain calculated through the latitude and longitude of the receiver at different times [82,92].

The camera data source includes any optical device that can capture digital images. It is assumed that these devices are relatively close to the earth's ground, which excludes satellites and other devices orbiting the Earth. Therefore, cameras are an affordable data source for both individuals and organizations. The size and weight of these devices can vary without drastically affecting the quality of the data produced, allowing them to be attached to and used in conjunction with various devices such as drones [62], smartphones [81], or ground robots [59]. An advantage of this type of data source is the high spatial resolution obtained from the captured images or videos. In [60], the authors used images with a spatial resolution of between 4 and 8 cm, captured by UAVs.

Satellites are an essential data source for obtaining data on sizeable agricultural land. A set of sensors attached to the satellite is used to capture the data. The data obtained are unified into satellite images, products composed of multiple spectral bands containing the information. Many products can be obtained from this data source, such as optical, SAR, or thermal images. Some essential characteristics of satellites, which differentiate them, are the type of product generated and their spatial and temporal resolutions. Of the resolutions, the former refers to the distance each image pixel represents on the ground. In contrast, the latter refers to the time it takes for each satellite to retrieve data from the exact location. From the articles analyzed, the use of six different satellites was identified: these are Google Earth [82], Sentinel-1 [58] and Sentinel-2 [82,91], Landsat 7, and Landsat 8 [63,92], MODIS [63].

The sensor group includes all IoT devices used statically in different locations to capture data. A wide variety of sensor types measure a single variable, such as temperature, radiation, rainfall, etc. [55]. Similarly, devices that include several sensors, such as a weather station or collars, are used on animals [75]. Using this data source usually requires taking care of the deployment, connection, and maintenance of the IoT devices. Some advantages of using sensors are that data obtained are particular to the area or task in which they are used and that they will be captured in real time and with a frequency to be defined by the user. On the other hand, the temporal resolution of sensors tends to be low, from seconds to minutes, so large amounts of information tend to be generated in a specific measurement period [76]. Wang and Mu explain that microsensors are being developed that are capable of capturing data on crop growth, land use, water use and characterization, and climate variables, among other essential aspects. The authors conclude that the use of these microsensors will enhance IoT development [70].

The “databases” group is different from the other categories as it compiles large data sets obtained from the other sources (sensors, images, surveys, etc.), which are made accessible through an interface defined by the entity. These databases allow easy and immediate access to a large amount of historical data, with accumulated records of up to dozens of years [86]. The vast majority of the databases identified are managed by public entities or government agencies, such as AWAP, CosmOZ, SILO, ASRIS, BOM, ISTAT, CNIR, IndiaStat, AAFC, ARPAS, ACIS, IMD, OGD, and KME. However, databases belonging to private entities were also identified, such as [95] using data sets from Brazilian cattle farms provided by the companies DSM Produtos Nutricionais Brasil S.A. and JBS S.A or [96] using a meteorological data set from Schneider Electric.

There is a common tendency among the analyzed articles to use more than one data source for research. This includes using data sources belonging to different categories (from those previously defined) and sources belonging to the same type but with different or complementary data to the other authorities.

4.3.1. Data Platforms in Agricultural Big Data

Of the articles analyzed, less than half detailed which platforms were used to store the collected data. From the reports that detailed this information, Figure 15 shows the distribution of uses of the mentioned platforms, categorized into Hadoop, Relational Database, NoSQL Database, and Cloud.

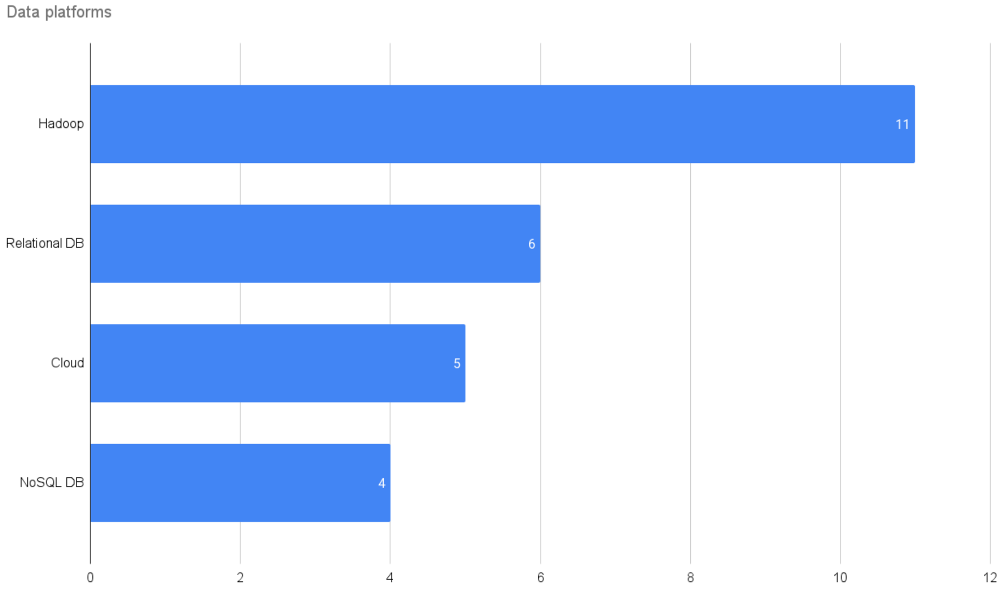


Figure 15. Data warehousing platforms used in Agricultural Big Data.

The Cloud category includes various cloud computing services such as AWS (Amazon Web Services) or GEE (Google Earth Engine). These services offer multiple benefits for operating with considerable amounts of data, such as Agricultural Big Data. A direct benefit of using these platforms is their large computational and storage capacities, which are suited to working with Big Data and can be resized according to the user's needs. Another benefit is the free and direct access they provide to different data sources, for example, access to satellite data captured by Landsat or Sentinel satellites.

Shelestov et al. [58] used AWS's quick and easy access to Sentinel 1 and 2 satellite imagery to work with datasets using up to 3 TB of memory space, eliminating the problems associated with downloading and storing data related to Big Data. Gumma et al. [92] list the following as reasons for using the GEE platform: the easy access to Landsat satellite data, the powerful computational capacity of the service, and the ability to perform parallel processing of the data, among others.

Of the analyzed articles, Wang et al. [77] use MongoDB, a document-oriented database, as temporary intermediate storage for data collected by sensors, which is later transferred to an implemented data warehouse. Sathiaraj et al. [83] used the REDIS in-memory database, whose data model is key-value, for the visualization system of the computational analyses performed because it has a low latency when accessing the data. Finally, Choudary et al. [54] used an RDF triplestore, a knowledge graph database, to store the data through a smart farm ontology. The database allows them to represent the relationships between two or more entities through the relationships in the graph.

The Relational DB group comprises RDBMSs (Relational Database Management Systems). Only three articles implemented relational databases in their research; of these, the RDBMSs identified are MySQL [80], PostgreSQL [75], Oracle Database, and the SQL server [77].

In Nobrega et al. [75], they use PostgreSQL to store the data obtained by the collars that were put on each sheep. They decided to use a relational database because their network of sensor collars has several entities, which can be efficiently designed in this type of database. They selected PostgreSQL from the available RDBMS options as it is suitable for environments working with the system's essential data, security, and integrity mechanisms. Yang et al. [80] use MySQL to persistently store data collected directly by sensors. The authors estimated that 86 million records would be stored in a month. They are aware that when you have tens of millions of records in the same table, the efficiency of data operations performed by MySQL decreases considerably. Therefore, they implemented two strategies to solve this problem; the first is to split the table with the millions of records into discrete tables daily. The second is to perform the data analysis processes in a separate service based on the Hadoop ecosystem.

The most identified category among the analyzed articles was Hadoop, which includes the data warehousing technologies of the Hadoop ecosystem, i.e., HDFS, Hive, and HBase. A fundamental difference between databases (Relational and NoSQL) with Hadoop technologies is that the latter is designed to store data in a distributed manner from the outset, which makes them highly suitable for working in the presence of Big Data.

Priya et al. [57] use HDFS to store the collected datasets due to its easy integration with the MapReduce paradigm, which enables parallel processing of the data, allowing them to work efficiently with large datasets. In Wang et al., HDFS is used as a single platform containing data from multiple relational and non-relational databases [77]. Data are synchronized from the different DBs to HDFS using technologies such as NiFi, Sqoop, or Flume. The data stored in HDFS are, in turn, filtered using Spark SQL to extract and store in Hive the data that will be used for analysis with machine learning techniques and in HBase the data that will be used directly for monitoring or visualization. The authors comment that with this data storage structure, they obtain scalability, high fault tolerance, and good performance for data processing.

4.3.2. Machine Learning Algorithms

Figure 16 shows all the ML techniques identified in the analyzed articles. A total of 36 different techniques were found, of which the five most used were Neural Networks, Random Forest, Support Vector Machines, Decision Trees, and Convolutional Neural Networks. From the graph, it can be seen that more than half of the techniques identified accumulate a single use and that the first seven techniques together accumulate a more significant benefit than the sum of the services of the rest of the methods.

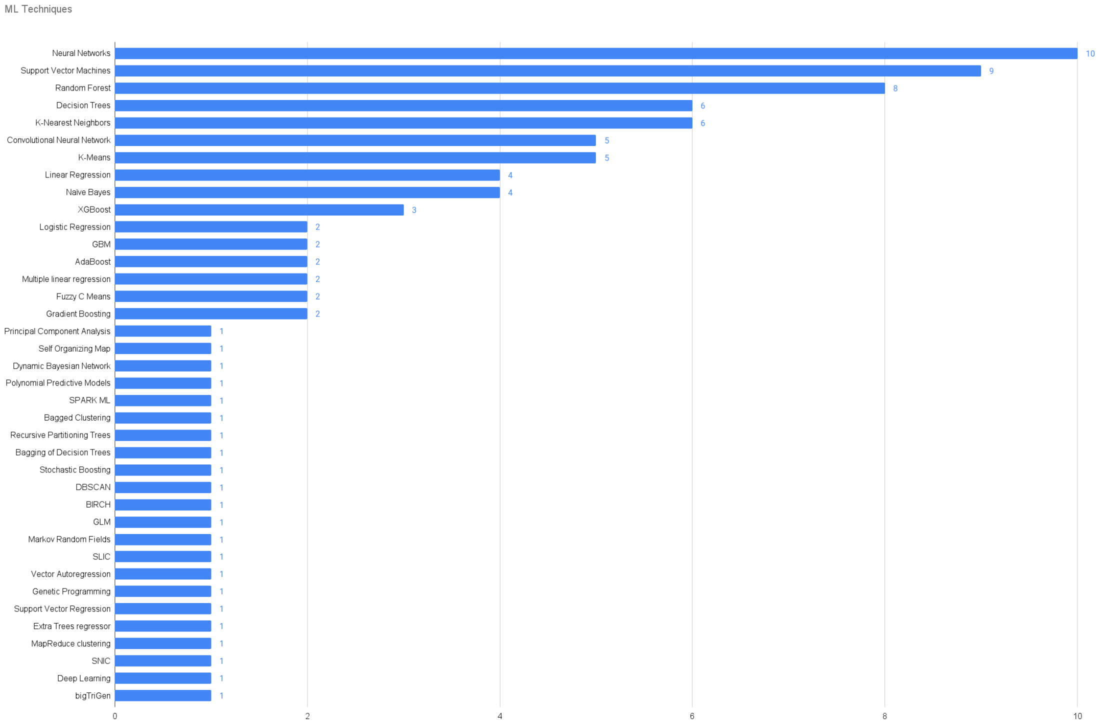


Figure 16. ML techniques in Agricultural Big Data.

The articles identified a common tendency to employ several ML techniques. This mainly occurs in two scenarios. The first arises because each algorithm processes the data differently, so the results obtained are influenced by the dataset and its characteristics (number of records, dimensionality, etc.). Therefore, several ML techniques are used, and the performances and results obtained by each model are compared. The second scenario is when they are used complementary for data analysis, i.e., the results obtained by one of the techniques are used for training or validation of the other techniques.

Balducci et al. [53] compare the implementation of diverse ML techniques against five tasks designed by them. For the tasks, they used other datasets, and the authors selected the compared techniques depending on the expected result. The techniques used are Neural Networks, Linear Regression, Polynomial Regression, Decision Tree, K-NN, and similarity clustering. Amani et al. [82] use the SNIC algorithm to segment layered stacked mosaic images. The segments obtained by the algorithm were subsequently used to perform an object-based classification using neural networks, which they used to generate an ACI (Annual Space-Based Crop Inventory) map of Canada.

There are several challenges related to using ML in the presence of Agricultural Big Data. As discussed earlier, the performance of ML algorithms is influenced by the dataset used. One of the challenges faced by Dutta et al. [63] is the high dimensionality that datasets tend to have when working with Big Data. The authors propose a domain-specific dimensionality reduction process, which allows them to increase both the efficiency of the system and the accuracy of ML algorithms.

Another challenge, associated only with some ML techniques (CNN, RNN, etc.), is a large amount of memory space they occupy to store the trained models and their parameters. For example, Tombe et al. [62] had to set aside the deep learning networks VGG16, VGG19, and GoogleNet due to their memory space limitations in the research. In these techniques, the space required will increase as more data are used for training.

Finally, the most recurrent challenge among the analyzed articles is the one associated with the quality of the data used for training ML models. Outliers, missing values, noise, and duplicate values can negatively impact the performance of an ML model, as it learns from these underlying relationships that do not exist in reality [88]. As a common challenge when using ML techniques, several authors use methods to improve the dataset's quality; these methods are known as data preprocessing.

In Doshi et al. [64], they perform preprocessing in two steps, the first is to replace missing values with large negative values that the models can treat as outliers, and the second step is to generate class labels using mathematical functions for later use in supervised learning techniques. On the other hand, Choudhary et al. [54] also perform a preprocessing of the dataset; in this case, columns with missing values are removed, and categorical data in String are encoded into integer values to make them usable by ML models. Aiken et al. [95] also process the data to standardize the two datasets used, transforming String data to lower case, removing special characters, expanding acronyms, and correcting common misspellings. The authors comment that the necessary standardizations are specific to each dataset.

Finally, in Melgar-García et al. [69], a new novel triclustering approach is presented, which is based on the evolutionary algorithm bigTriGen, with which three-dimensional patterns are extracted. Venkatesan et al.[73] use a number of ML algorithms, including NN, support vector regression, random forest, K-nearest neighbors, extreme gradient boosting and gradient boosting machine, and time series algorithm ARIMA with binary classification for a different number of input features.

5. Discussion

The various solutions for Agricultural Big Data include analyzing structured, semi-structured, and unstructured data. For example, sensors and satellites are available through databases of structured data. Despite their availability, these data often need to be processed due to various problems such as (1) incomplete data records, (2) the data need to be read from a raw file, and (3) validation of the data before it can be used. On the other hand, semi-structured data, such as data stored in No-SQL databases, require processing to achieve a structure that allows analysis through techniques such as ML.

Unstructured data present the most significant challenges, as they require the use of systems that allow for reading, data extraction, validation, data transformation (from one type to another, or the creation of new data), and loading onto a platform that is capable of storing large volumes of data. For example, we found unstructured data from satellites, drones, cameras, and videos in the selected papers.

In recent years, with the rise of satellite data, study on crop development status monitoring (including crop yield estimation, crop classification, and crop leaf area) has slowly shifted from large-scale stationary range information return to cooperative research of mesoscale, near real-time, multisource data; this shift has presented higher needs for remote sensing data purchase, and analysis efficiency [97].

According to Xu et al. [98], remote sensing is an essential non-physical Earth observation method widely applied in agriculture, water, climate, and other areas. On the other hand, the progress of IoT and computing technologies has enabled significant advances in remote sensing. Today, more than 1000 remote sensing satellites have been launched, from which data are accumulating at a rate of terabytes per day. In addition, many other approaches, such as uncrewed aerial vehicles, are frequently used for remote sensing data collection. Based on multivariate remote sensing data, Earth's Big Data has reached the ZB level.

Therefore, several challenges are to be faced when using this data type. Firstly, the amount of remotely sensed data expected to be managed is enormous, and the structure of remotely sensed data is complex. Moreover, remote sensing data are stored in different formats and systems, such as GeoTiff, ASCII, and HDF, which are incompatible. Secondly, it is noted that remote sensing data processing places high demands on computational performance, which requires the effort of continuous improvement and data accuracy; in addition to the development of ML algorithms, algorithms for remote sensing data processing are becoming more complex [98].

To ensure the high availability of both raw and structured data, distributed storage systems have been widely applied. For example, several works have used MongoDB, a distributed database that supports storing and indexing remotely sensed data, vector data, and semi-structured data. On the other hand, Hadoop's distributed file system can be applied to store all types of data, from structured to remote sensing data, making it a perfect solution for volume and variety of data. Cloud Computing technology also improves the availability of satellite data and reduces the development time of systems for Agricultural Big Data. It is then possible to process the data for analysis without installing servers containing specialized tools. In the future, the use of Cloud Computing will increase, thanks to the scalability, cost reduction, and availability of several ML classifiers (such as CART, Random Forest, SVM, and ANN). However, developing and applying complex ML models to large geospatial datasets is not trivial and may require leveraging multiple computing platforms or tools.

GEE provides access to large-scale remote sensing data sets and computing resources. However, even though GEE is a public platform, it does not meet all the requirements of data scientists [98]. Some of the problems are the lack of flexibility in a large amount of stored data, the scalability in processing on demand, the problematic use of MapReduce, the integrity problems of the Big Data tools with those of remote sensing, and the algorithms of ML that are still basic concerning the needs of analysis [98]. For example, the computational model currently employed by GEE does not work well for recursion processes and operations that require a large amount of cached data, such as training many ML models. One solution is to train models at scale on another private platform and use GEE for data management, pre-processing, image classification, post-processing, and visualization.

This will bring a greater volume of data, as well as more significant processing challenges. Data Lakes, a type of repository that allows structured and unstructured data to be stored in different storage areas [14], will enable agricultural scientists to understand the relationships between data and the types of analysis that can be implemented. This is because the Data Lake has a model representing metadata characteristics and the possible relationships between the data [99].

According to Batini et al. [100], data quality is influenced by the type of data to be analyzed, their sources, and their application domains. He explains that the data life cycle must be examined for each case since there are flaws at each stage. For data obtained from sensors, weakness comes from the capacity of a network to store all data available from sensors, in addition to the location of data, as this is often inaccurate. On the other hand, it is necessary to identify the needed data since some are no longer used, influencing their life cycle.

Data Lakes allow data management from two points of view [99]. The first is related to the characteristics of each data type, whether structured or non-structured. The second has to do with the traceability of the data since they can be stored in more than one repository area when processed. Flexible Data Lake architecture enables faster data loading and parallel processing, resulting in faster instant analytical insight. In addition, data management improves data quality since more information is available on the data's life cycle due to the availability of metadata and the relationships between the data.

6. Limitations of the Study

This section describes four types of threats to the validity of this review.

6.1. Construct Validity

The study of this threat is relevant for ranking the selected studies [101]. On the one hand, keywords used for the search were proposed by one of the authors and validated by the others. The results are the terms Big Data, ML, and agriculture. However, the list of terms is incomplete; some alternative terms may alter the list of selected papers. On the other hand, the search string was used in IEEE Xplore, ACM, WoS, Springer, MDPI, and Scopus. We found most of the research papers on agricultural Big Data and ML through the search string terms. To mitigate the threat of not considering essential reports, we explored corresponding articles from review studies and state-of-the-art research.

6.2. Internal Validity

This type of validity is concerned with the study of analyzing the data extracted from the selected papers [102]. A research assistant collected the data and then classified them according to key concepts given by the group of researchers. There was a reasonable level of agreement between authors, with a kappa coefficient value of 0.9. This significantly reduces the threats of dissimilarity by showing a similar understanding of relevance.

6.3. External Validity

External validity corresponds to the context of the investigation [102]. Furthermore, the results have been evaluated in Agricultural Big Data and ML. Therefore, the terms utilized in the search string and the classification of the articles can help professionals in the agricultural area and the use of data in Big Data and ML.

6.4. Conclusion Validity

This threat is related to identifying inappropriate relationships that lead to an incorrect conclusion. On the other hand, this threat to the validity of the conclusion refers to different elements of analysis, such as incorrect data extraction, review of missing studies, and identification of incorrect gaps, among others. To reduce this threat, a PRISMA-based criterion has been defined for the data extraction and selection process [102].

7. Conclusions

The study's objective was to identify the types of data used in Agricultural Big Data and ML, according to their condition, whether structured, semi-structured, or unstructured. Furthermore, the data types were classified according to the data sources, such as repositories, databases, and platforms, identifying new data types and trends. The methodology used was the Systematic Literature Review according to the PRISMA criteria and Kitchemham and Chartes protocol. The results obtained are a basis for future research in the data application in Agricultural Big Data and ML.

Forty-three papers were analyzed. The primarily structured data used were climate characteristics such as temperature, humidity, atmospheric pressure, and wind speed; soil characteristics such as soil type, pH, and salinity, among others; and crop data such as plant characteristics, soil characteristics, and information from databases or repositories. The data were classified as Int, Float, String, Date, and Time. In addition, microsensors are beginning to be used to monitor crops from different points of view, such as water use, soil data, climate, and data from the plant itself.

On the other hand, semi-structured data came from databases, files, and No-SQL databases. Some data were stored in JSON, key-value, RDF, and spreadsheet formats. These data need to be processed to store in structured databases that allow for easy access and analysis by ML tools.

The unstructured data came from satellite images, GPS, video cameras, and documents. The most significant amount of data comes from satellite imagery, as there are more than 1000 available, with various sensors measuring different terrain characteristics, such as LiDAR and RGB.

The main storage platforms are structured databases, No-SQL databases, Hadoop, and Cloud Computing. Cloud Computing is expected to increase the capacity of remote sensing analysis through new ML algorithms for processing and analysis. This will allow researchers to speed up the implementation of Agricultural Big Data.

For its part, GEE allows a large amount of remotely sensed data to be processed and stored. However, the processing and analysis of complex data are still limited. In the future, more features and data will become available, which will help researchers to generate more complete and accurate reports for decision making.

Data Lakes are still underutilized despite advances in using Big Data in agriculture. This brings enormous challenges since the architecture must be modeled considering the layers to be used for the different types of data. In addition, the data must be stored and managed through metadata catalogs. Unfortunately, none of the analyzed papers mention using such data management tools. However, we believe that in the coming years, these Data Lakes will be used in the application of Agricultural Big Data due to the need for the different repositories available so far to become integrated.

To create an optimal strategy for data storage and processing and an appropriate software selection, a skilled team with a spectrum of knowledge and skills in data management, modern data architecture, Big Data technologies, as well as experts in the problem's domain, will be necessary [14]. This study provides information to improve data management by: (1) classifying the types of data used in Agricultural Big Data and ML, (2) providing information on data structures and sources, and (3) providing information on storage platforms and trends.

Author Contributions: A.C. contributed with the direction of the SRL, the organization of the work, and the writing of the paper. S.P. contributed with student work, writing, and formatting. P.G. and J.L.F. contributed to the methodological flow and discussion. M.C. contributed figures, tables, and spelling and grammar checking. All authors have read and accepted the published version of the manuscript.

Funding: Universidad de La Frontera, Project PAT22-0005(2023).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: ANID project code FOVI210012.

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

Appendix A

Table A1. Summary of data by paper.

Title	Category	Unstructured Data	Structured Data
Fruit disease prediction using machine learning over big data	Camera	camera images	stage of diseased fruit.
A hybrid machine learning approach to automatic plant phenotyping for smart agriculture	Camera	camera images	image hue, pixels.
A framework for the management of agricultural resources with automated aerial imagery detection	Camera, Database	drone images, digital surface models	type of tree, x (pixel), y (pixel), size, confidence, land, cover, class.
AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms	Database	X	soil type, aquifer thickness, soil pH, thickness of topsoil, precipitation, temperature, latitude, longitude, distance from sea, rainfall, area under cultivation.
Record linkage for farm-level data analytics: Comparison of deterministic, stochastic, and machine learning methods	Database	X	state, city, name, state ID, owner's name.
Reference evapotranspiration estimation and modeling of the Punjab Northern India using deep learning	Database	X	max air temperature, min air temperature, average temperature, relative humidity, wind speed, solar radiation, sunshine hours, reference evotranspiration.
Big data and machine learning for crop protection	Database		soil pH, state of Shire, winter crop, amount of cultivation, stubble management.
Adaptive and Efficient Streaming Time Series Forecasting with Lambda Architecture and Spark	Database	X	X
Towards modeling beef cattle management with Genetic Programming	Database	X	Consistency of all animals, percentage of artificial insemination, mean value of age of cows expressed in days, mean number of parturitions per cow, number of occurred deliveries, mean number of necessary inseminations which resulted in positive pregnancy diagnosis, etc.
Enhancement of Convolutional Neural Networks Classifier Performance in the Classification of IoT Big Data	Database	X	X
An adaptive model for forecasting seasonal rainfall using predictive analytics	Database	X	min temperature, max temperature, avg temperature, humidity, wind speed, precipitation, wind direction, cloud cover, visibility, atmosphere pressure.
Computer Vision for Smart Farming and Sustainable Agriculture	Database, Camera	drone images	X
Application of Google Earth Engine cloud computing platform, Sentinel imagery, and neural networks for crop mapping in Canada	Database, GPS, Satellite	SAR images, optical images	cropland class, number of field samples, VV and VH, NDVI and NDWI.
Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement	Database, Sensors	X	crop type, year of the time series, province, altitude total area, cultivation area, total crop production, total harvest production, temperature average, min and max, rainfall amount, amount of phosphate and potash minerals, etc.
Predicting climate types for the Continental United States using unsupervised clustering techniques	Database, Sensors	X	max temp, min temp, precipitation, month, year.
YieldPredict: A Crop Yield Prediction Framework for Smart Farms	Database, Sensors	X	State Name, district name, crop year, season, crop, area, production, seasonal rainfall, nitrogen, phosphorus, potassium, saline soil (ha), sodic soil (ha).
Cloud Approach to Automated Crop Classification Using Sentinel-1 Imagery	Satellite	SAR images	VV and VH polarizations.
Scalable distributed random forest classification for paddy rice mapping Asian conference on remote sensing ACRS 2019	Satellite, Camera, Database	optical images, SAR images	VV and VH coefficients, NDVI, NDWI, PSRI, B8A, B12, B11, B08, B07, B06, B05, B04, B03, B02.
Real-Time Big Data Analytics for Agricultural Land Hotspot Prediction	Satellite, database	satellite images, CSV	precipitation, min temperature, max temperature, avg temperature, cloud cover, vapor pressure, wet day frequency, ground frost frequency, color bands ->RGB.
Big Data Architecture for Environmental Analytics	Satellite, Database, Person	satellite images, Domain Knowledge	X
Crop Prediction on the Region Belts of India: A Naïve Bayes MapReduce Precision Agricultural Model	Satellite, Database, Sensors	satellite images	soil moisture, rainfall, temperature, atmospheric pressure, relative humidity, wind speed, wind direction, soil temperature, radiation, diffusion rate.

Table A1. *Cont.*

Title	Category	Unstructured Data	Structured Data
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	Satellite, GPS	satellite images	RGB bands, NIR band, SWIR 1 and SWIR 2 bands, thermal bands, NDVI index, EVI index, NDWI index, latitude, longitude, slope.
Evaluation of agricultural climate and regional agricultural economic efficiency based on remote sensing analysis	Satellite, GPS	satellite images	latitude, longitude, position, shape, color, average, temperature, total power of agricultural development machinery, fixed employees in agriculture, forestry, animal husbandry and fishery, area of land used at the end of each year, amount of fertilizer, plastic used.
Animal monitoring based on IoT technologies	Sensors	camera video	hours of activity, travel times, preferred pasture areas and timings, anomalous situations, number of fence and posture infractions.
Machine learning prediction analysis using IoT for smart farming	Sensors	text, Web data, CSV	temperature, humidity.
Big data analytics and artificial intelligence serving agriculture	Sensors	X	temperature, dew point, humidity, pressure, visibility, wind direction, wind speed, burst speed, events, rain, weather conditions, month, day.
The Implementation of a Practical Agricultural Big Data System	Sensors	X	air temperature, air humidity, light intensity, soil moisture, fruit size, branch length, soil salinity, wind speed, wind direction, conductivity, soil temperature, CO2, pH.
Architecture Design of a Smart Farm System Based on Big Data Appliance Machine Learning	Sensors	X	color, weight, inner temperature, outer temperature, humidity.
Intelligent computational techniques for crops yield prediction and fertilizer management over big data environment	Sensors	X	temperature, rainfall, pH, water level, nitrogen, phosphorous, potassium, calcium, magnesium, sulfur.
Big data in precision agriculture: Weather forecasting for future farming.	Sensors	X	min temperature, max temperature, humidity, rainfall.
Botanical Internet of Things: Toward Smart Indoor Farming by Connecting People, Plant, Data and Clouds	Sensors, Camera	camera video	temperature, humidity, illumination intensity, air gasses (CO2, O2, O3, NO2), plant ID, time.
The effective yield of paddy crop in Sivaganga district – An initiative for smart farming	Sensors, Database	X	crop yield, farmer's name, Place–Village–Panchayat–Taluk–District, survey number, contact details, soil type, paddy variety, types of fertilizers used at the different stages of cultivation, field size, Irrigation type.
An application of machine learning technique in forecasting crop disease	Sensors, Database, Person	X	temperature, relative humidity, wind speed, wind direction, precipitation accumulation, solar radiation, potato blight disease.
Design of Smart Agriculture Systems using Artificial Intelligence and Big Data Analytics	Camera	images, video, audio, telemetry data, user data,	temperature, light, humidity, and soil moisture.
Innovation of agricultural economic management in the process of constructing smart agriculture by big data	Database	GIS Data, social network, GPS Data	egg price, duck egg price, export volume, output, market elasticity, labor force change and inventory as variables.
A new big data triclustering approach for extracting three-dimensional patterns in precision agriculture	Satellite	Images	three-dimensional datasets, NDVI index, Soil-Adjusted Vegetation Index (SAVI) and the Enhanced Vegetation Index (EVI), and GNDVI.
Agricultural Irrigation Recommendation and Alert (AIRA) system using optimization and machine learning in Hadoop for sustainable agriculture	Sensors	X	temperature (°C), relative humidity (%), mean sea level pressure (hPa), snowfall amount (cm), sunshine duration (Min), evapotranspiration (Mm), FAO reference evapotranspiration (mm), wind speed (Km/hr), wind direction (A°), soil temperature (A°C), soil moisture (fraction).
Risk monitoring model of intelligent agriculture Internet of Things based on big data	Sensors	X	X
Application of Modern GIS and Remote Sensing Technology Based on Big Data Analysis in Intelligent Agriculture	Satellite	Images, document data.	X
Superior fuzzy enumeration crop prediction algorithm for big data agriculture applications	Database	X	Crop field dataset, name area.
Big Data Scheme from Remote Sensing Applications: Concluding Notes for Agriculture and Forestry Applications	Satellite	X	X
A Machine Learning Based Model for Energy Usage Peak Prediction in Smart Farms.	Sensors	X	Internal temperature, internal humidity, ventilation temperature, heating temperature, outside temperature, outside solar temperature, dew point, hourly accumulated light, hourly solar radiation, temperature difference, crop output production.
Monitoring Complex Integrated Crop–Livestock Systems at Regional Scale in Brazil: A Big Earth Observation Data Approach.	Satellite	X	NDVI, EVI, Near-Infrared spectral band (NIR) and Mid-Infrared spectral band (MIR).

References

1. Praveen, B.; Sharma, P. A review of literature on climate change and its impacts on agriculture productivity. *J. Public Aff.* **2019**, *19*, e1960. [CrossRef]
2. Yaqoob, N.; Ali, S.A.; Kannaiah, D.; Khan, N.; Shabbir, M.S.; Bilal, K.; Tabash, M.I. The effects of Agriculture Productivity, Land Intensification, on Sustainable Economic Growth: A panel analysis from Bangladesh, India, and Pakistan Economies. *Environ. Sci. Pollut. Res. Int.* **2022**, 1–9.. [CrossRef]
3. Wakelin, S.A.; Gomez-Gallego, M.; Jones, E.; Smaill, S.; Lear, G.; Lambie, S. Climate change induced drought impacts on plant diseases in New Zealand. *Australas. Plant Pathol.* **2018**, *47*, 101–114. [CrossRef]
4. Liakos, K.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine Learning in Agriculture: A Review. *Sens. Multidiscip. Digit. Publ. Inst.* **2018**, *18*, 2674. [CrossRef]
5. Sundmaeker, H.; Verdouw, C.; Wolfert, S.; Pérez Freire, L. Internet of Food and Farm. In *Digitising the Industry-Internet of Things Connecting the Physical, Digital and Virtual Worlds*; Vermesan, O., Friess, P., Eds.; River Publishers: Delft, Denmark, 2017.
6. Wolfert, S.; Ge, L.; Verdouw, C.; Bogaardt, M.J. Big data in smart farming–A review. *Agric. Syst.* **2017**, *153*, 69–80. [CrossRef]
7. Nandyala, C.; Kim, H.K. Big and meta data management for U-agriculture mobile services. *Int. J. Software Eng. Appl. IJSEIA* **2016**, *10*, 257–270. [CrossRef]
8. Cravero, A.; Sepúlveda, S. Use and Adaptations of Machine Learning in Big Data—Applications in Real Cases in Agriculture. *Electronics* **2021**, *10*, 552. [CrossRef]
9. Ihde, N.; Marten, P.; Eleliemy, A.; Poerwawinata, G.; Silva, P.; Tolovski, I.; Ciorba, F.M.; Rabl, T. A Survey of Big Data, High Performance Computing, and Machine Learning Benchmarks. In *Proceedings of the Technology Conference on Performance Evaluation and Benchmarking*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 98–118.
10. Wu, Y.; Xiang, Y.; Ge, J.; Muller, P. High-performance computing for big data processing. *Future Gener. Comput. Syst.* **2018**, *88*, 693–695. [CrossRef]
11. Sun, A.; Scanlon, B. How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environ. Res. Lett. IOP Publ.* **2019**, *14*, 73001. [CrossRef]
12. Cravero, A.; Pardo, S.; Sepúlveda, S.; Muñoz, L. Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review. *Agronomy* **2022**, *12*, 748. [CrossRef]
13. Saiz-Rubio, V.; Rovira-Más, F. From smart farming towards agriculture 5.0: A review on crop data management. *Agronomy* **2020**, *10*, 207. [CrossRef]
14. Šuman, S.; Poščić, P.; Gligora Marković, M. Big Data Management Challenges. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 717–723. [CrossRef]
15. Bhatnagar, R. Machine learning and big data processing: A technological perspective and review. In *Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 468–478.
16. Rasyid, L.; Andayani, S. Review on clustering algorithms based on data type: Towards the method for data combined of numeric-fuzzy linguistics. In *Proceedings of the 5th International Conference on Research, Implementation, & Education of Mathematics and Sciences, 7–8 May 2018, Yogyakarta, Indonesia*; IOP Publishing: Bristol, UK, 2018; Volume 1097, p. 012082.
17. Nandi, G.; Sharma, R.K. *Data Science Fundamentals and Practical Approaches: Understand Why Data Science Is the Next*; BPB Publications: Uttar Pradesh, India, 2020.
18. Firdaus, H.; Hassan, S.I. Unsupervised Learning on Healthcare Survey Data with Particle Swarm Optimization. In *Machine Learning with Health Care Perspective*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 57–89.
19. Kitchenham, B.; Charters, S. Guidelines for performing systematic literature reviews in software engineering. Thechnical Rep. Ebse´07. 2007. Available online: https://www.researchgate.net/publication/302924724_Guidelines_for_performing_Systematic_Literature_Reviews_in_Software_Engineering (accessed on 25 September 2022).
20. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Ann. Intern. Med.* **2009**, *151*, 264–269. [CrossRef] [PubMed]
21. Lin, C.S.; Wang, Y.Y. Data type and data source preferences for six social sciences subjects in quantitative data reuses. *Proc. Assoc. Inf. Sci. Technol.* **2018**, *55*, 867–868. [CrossRef]
22. Putra, H.Y.; Putra, H.; Kurniawan, N.B. Big data analytics algorithm, data type and tools in smart city: A systematic literature review. In Proceedings of the 2018 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, Indonesia, 22–26 October 2018; pp. 474–478.
23. Fassnacht, F.; Hartig, F.; Latifi, H.; Berger, C.; Hernández, J.; Corvalán, P.; Koch, B. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sens. Environ.* **2014**, *154*, 102–114. [CrossRef]
24. Roy, D.; Shirazi, F. A Review on Multiple Data Source Based Recommendation Systems. In Proceedings of the 2021 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 15–17 December 2021; pp. 1534–1539.
25. Sassi, I.; Ouaftouh, S.; Anter, S. Adaptation of Classical Machine Learning Algorithms to Big Data Context: Problems and Challenges. In Proceedings of the 2019 1st International Conference on Smart Systems and Data Science (ICSSD), Rabat, Morocco, 3–4 October 2019; pp. 1–7.

26. Elshawi, R.; Sakr, S.; Talia, D.; Trunfio, P. Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service. *Big Data Res.* **2018**, *14*, 1–11. [CrossRef]
27. Haig, B.D. Big data science: A philosophy of science perspective. In *Big Data in Psychological Research*; Woo, S.E., Tay, L., Proctor, R.W., Eds.; American Psychological Association: Washington, DC, USA, 2020; pp. 15–33.
28. Santos, M.; e Sá, J.; Costa, C.; Galváo, J.; Andrade, C.; Martinho, B.; Lima, F.; Costa, E.; Lima, F. A big data analytics architecture for industry 4.0. In *Proceedings of the World Conference on Information Systems and Technologies, Madeira, Portugal, 11–13 April 2017*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 175–184.
29. Salma, C.A.; Tekinerdogan, B.; Athanasiadis, I.N. *Chapter 4—Domain-Driven Design of Big Data Systems Based on a Reference Architecture*; Morgan Kaufmann: Burlington, MA, USA, 2017; pp. 49–68. [CrossRef]
30. Sowmya, R.; Suneetha, K. Data mining with big data. In Proceedings of the 2017 11th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, India, 5–6 January, 2017; pp. 246–250.
31. Song, I.Y.; Zhu, Y. Big data and data science: What should we teach? *Expert Syst.* **2016**, *33*, 364–373. [CrossRef]
32. Demchenko, Y.; De-Laat, C.; Membrey, P. Defining architecture components of the big data ecosystem. In Proceedings of the 2014 International Conference on Collaboration Technologies and Systems, CTS 2014, Minneapolis, MN, USA, 19–23 May 2014; pp. 104–112.
33. Semlali, B.E.B.; Amrani, C.E.; Ortiz, G. Hadoop paradigm for satellite environmental big data processing. *Int. J. Agric. Environ. Inf. Syst.* **2020**, *11*, 23–47. [CrossRef]
34. Alex, S.A.; Kanavalli, A. Intelligent computational techniques for crops yield prediction and fertilizer management over big data environment. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 3521–3526. [CrossRef]
35. Cherkassky, V.; Mulier, F. *Learning from Data: Concepts, Theory, and Methods*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
36. Rudin, C.; Wagstaff, K. Machine learning for science and society. *Mach Learn.* **2014**, *95*, 1–9. [CrossRef]
37. Qiu, J.; Wu, Q.; Ding, G.; Xu, Y.; Feng, S. A survey of machine learning for big data processing. *Eurasip J. Adv. Signal Process.* **2016**, *1*, 1–16.
38. Benos, L.; Tagarakis, A.C.; Dolias, G.; Berruto, R.; Kateris, D.; Bochtis, D. Machine Learning in Agriculture: A Comprehensive Updated Review. *Sensors* **2021**, *21*, 3758. [CrossRef] [PubMed]
39. Bal, S.K. Agro-meteorological basis of extremes of temperature with special perspective to livestock and poultry. *Clim. Resilient Anim. Husb.* **2021**, 23.
40. Malik, A.; Burney, A.; Ahmed, F. A comparative study of unstructured data with SQL and NO-SQL database management systems. *J. Comput. Commun.* **2020**, *8*, 59–71. [CrossRef]
41. Villars, R.L.; Olofson, C.W.; Eastwood, M. *Big Data: What It Is and Why You Should Care*; White Paper: Framingham, MA, USA, 2011.
42. Eberendu, A.C.; Madonna University. Unstructured Data: An overview of the data of Big Data. *Int. J. Comput. Trends Technol.* **2016**, *38*, 46–50. [CrossRef]
43. Sánchez, M.; Barrena, M.; Bustos, P.; Campillo, C.; García, P. Arquitectura software basada en tecnologías smart para agricultura de precisión. *Jornadas Ing. Softw. Bases Datos* **2020**, *219*, 219–349.
44. Sambrekar, K.; Rajpurohit, V.S.; Joshi, J. A proposed technique for conversion of unstructured Agro-data to semi-structured or structured data. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018.
45. Alkathiri, M.; Jhummarwala, A.; Potdar, M.B. Multi-dimensional geospatial data mining in a distributed environment using MapReduce. *J. Big Data* **2019**, *6*, 1–34. [CrossRef]
46. Guimarães, N.; Pádua, L.; Marques, P.; Silva, N.; Peres, E.; Sousa, J.J. Forestry Remote Sensing from Unmanned Aerial Vehicles: A review focusing on the data, processing and potentialities. *Remote Sens.* **2020**, *12*, 1046. [CrossRef]
47. Press, F.; Siever, R. Earth, 1998. Available online: https://aws.amazon.com/earth/ (accessed on 19 November 2022).
48. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
49. Pekel, J.F.; Cottam, A.; Gorelick, N.; Belward, A.S. High-resolution mapping of global surface water and its long-term changes. *Nature* **2016**, *540*, 418–422. [CrossRef]
50. Padarian, J.; Minasny, B.; McBratney, A.B. Using Google's cloud-based platform for digital soil mapping. *Comput. Geosci.* **2015**, *83*, 80–88. [CrossRef]
51. Landset, S.; Khoshgoftaar, T.M.; Richter, A.N.; Hasanin, T. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *J. Big Data* **2015**, *2*, 1–36. [CrossRef]
52. Odun-Ayo, I.; Ananya, M.; Agono, F.; Goddy-Worlu, R. Cloud computing architecture: A critical analysis. In Proceedings of the 2018 18th International Conference on Computational Science and Applications (ICCSA), Melbourne, Australia, 2–5 July 2018.
53. Balducci, F.; Impedovo, D.; Pirlo, G. Machine learning applications on agricultural datasets for smart farm enhancement. *Machines* **2018**, *6*, 38. [CrossRef]
54. Choudhary, N.K.; Chukkapalli, S.S.L.; Mittal, S.; Gupta, M.; Abdelsalam, M.; Joshi, A. YieldPredict: A Crop Yield Prediction Framework for Smart Farms. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 2340–2349. [CrossRef]

55. Gnanasankaran, N.; Ramaraj, E. The Effective Yield Of Paddy Crop In Sivaganga District-An Initiative For Smart Farming. *Int. J. Sci. Technol. Res.* **2020**, *9*, 2.
56. Donzia, S.K.Y.; Kim, H.k. Architecture Design of a Smart Farm System Based on Big Data Appliance Machine Learning. In Proceedings of the 2020 20th International Conference on Computational Science and Its Applications (ICCSA), Cagliari, Italy, 1–4 July 2020; pp. 45–52. [CrossRef]
57. Priya, R.; Ramesh, D.; Khosla, E. Crop Prediction on the Region Belts of India: A Naïve Bayes MapReduce Precision Agricultural Model. In Proceedings of the 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, 19–22 September 2018; pp. 99–104.
58. Shelestov, A.; Lavreniuk, M.; Vasiliev, V.; Shumilo, L.; Kolotii, A.; Yailymov, B.; Kussul, N.; Yailymova, H. Cloud Approach to Automated Crop Classification Using Sentinel-1 Imagery. *IEEE Trans. Big Data* **2019**, *6*, 572–582. [CrossRef]
59. Yahata, S.; Onishi, T.; Yamaguchi, K.; Ozawa, S.; Kitazono, J.; Ohkawa, T.; Yoshida, T.; Murakami, N.; Tsuji, H. A hybrid machine learning approach to automatic plant phenotyping for smart agriculture. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 4–19 May 2017; pp. 1787–1793.
60. Ochoa, K.S.; Guo, Z. A framework for the management of agricultural resources with automated aerial imagery detection. *Comput. Electron. Agric.* **2019**, *162*, 53–69. [CrossRef]
61. Fenu, G.; Malloci, F.M. An application of machine learning technique in forecasting crop disease. In *Association for Computing Machinery*; 2019; pp. 76–82. [CrossRef]
62. TOMBE, R. Computer Vision for Smart Farming and Sustainable Agriculture. In Proceedings of the 2020 IST-Africa Conference (IST-Africa), Kampala, Uganda, 18–22 May 2020.
63. Dutta, R.; Li, C.; Smith, D.; Das, A.; Aryal, J. Big Data Architecture for Environmental Analytics. *Int. Symp. Environ. Softw. Syst.* **2015**, 578–588.
64. Doshi, Z.; Nadkarni, S.; Agrawal, R.; Shah, N. AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018; pp. 1–6.
65. Rehman, A.; Liu, J.; Keqiu, L.; Mateen, A.; Yasin, M.Q. Machine learning prediction analysis using IoT for smart farming. *Int. J. Emerg. Trends Eng. Res.* **2020**, *8*, 6482–6487.
66. Tarik, H.; Mohammed, O.J. Big Data Analytics and Artificial Intelligence Serving Agriculture. In *Proceedings of the Advanced Intelligent Systems for Sustainable Development (AI2SD'2019), Marrakech, Morocco, 8–11 July 2019*; Ezziyyani, M., Ed.; Springer International Publishing: Cham, Switzerland, 2020; pp. 57–65.
67. Kumari, M.; Kumar, A.; Singh, P.; Singh, S. Multidisciplinary Real-Time Model for Smart Agriculture based on Weather Forecasting Using IoT, Machine Learning, Big Data and Cloud. In Proceedings of the 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 22–23 October 2021; pp. 1–5.
68. Katyayan, A.; Mashelkar, S.; DC, A.G.; Morajkar, S. Design of Smart Agriculture Systems using Artificial Intelligence and Big Data Analytics. In Proceedings of the 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 17–18 December 2021; pp. 592–595.
69. Melgar-García, L.; Gutiérrez-Avilés, D.; Godinho, M.T.; Espada, R.; Brito, I.S.; Martínez-Álvarez, F.; Troncoso, A.; Rubio-Escudero, C. A new big data triclustering approach for extracting three-dimensional patterns in precision agriculture. *Neurocomputing* **2022**, *500*, 268–278. [CrossRef]
70. Wang, Q.; Mu, Z. Risk monitoring model of intelligent agriculture Internet of Things based on big data. *Sustain. Energy Technol. Assess.* **2022**, *53*, 102654. [CrossRef]
71. Wang, X.; Yu, S.; Wen, Z.; Zhang, L.; Fang, C.; Jiang, L. Application of Modern GIS and Remote Sensing Technology Based on Big Data Analysis in Intelligent Agriculture. *J. Indian Soc. Remote. Sens.* **2022**, 1–11.. [CrossRef]
72. Ahamed, T. Big Data Scheme from Remote Sensing Applications: Concluding Notes for Agriculture and Forestry Applications. In *Remote Sensing Application*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 351–361.
73. Venkatesan, S.; Lim, J.; Ko, H.; Cho, Y. A Machine Learning Based Model for Energy Usage Peak Prediction in Smart Farms. *Electronics* **2022**, *11*, 218. [CrossRef]
74. Kuchler, P.C.; Simões, M.; Ferraz, R.; Arvor, D.; de Almeida Machado, P.L.O.; Rosa, M.; Gaetano, R.; Bégué, A. Monitoring Complex Integrated Crop–Livestock Systems at Regional Scale in Brazil: A Big Earth Observation Data Approach. *Remote Sens.* **2022**, *14*, 1648. [CrossRef]
75. Nóbrega, L.; Tavares, A.; Cardoso, A.; Gonzalves, P. Animal monitoring based on IoT technologies. In Proceedings of the 2018 IoT Vertical and Topical Summit on Agriculture-Tuscany (IOT Tuscany), Tuscany, Italy, 8–9 May 2018; pp. 1–5.
76. Yang, J.; Liu, M.; Lu, J.; Miao, Y.; Hossain, M.A.; Alhamid, M.F. Botanical Internet of Things: Toward Smart Indoor Farming by Connecting People, Plant, Data and Clouds. *Mob. Netw. Appl.* **2018**, *23*, 188–202. [CrossRef]
77. Wang, X.; Yang, K.; Liu, T. The Implementation of a Practical Agricultural Big Data System. In Proceedings of the 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, 6–9 December 2019; pp. 1955–1959.
78. Veerachamy, R.; Ramar, R. Agricultural Irrigation Recommendation and Alert (AIRA) system using optimization and machine learning in Hadoop for sustainable agriculture. *Environ. Sci. Pollut. Res.* **2022**, *29*, 19955–19974. [CrossRef] [PubMed]

79. Bendre, M.R.; Thool, R.C.; Thool, V.R. Big data in precision agriculture: Weather forecasting for future farming. In Proceedings of the 2015 1st International Conference on Next Generation Computing Technologies (NGCT) IEEE, Dehradun, India, 4–5 September 2015; pp. 744–750.
80. Yang, C.; Huang, Q.; Li, Z.; Liu, K.; Hu, F. Big Data and cloud computing: Innovation opportunities and challenges. *Int. J. Digit. Earth* **2017**, *10*, 13–53. [CrossRef]
81. Vasumathi, M.T.; Kamarasan, M. Fruit disease prediction using machine learning over big data. *Int. J. Recent Technol. Eng.* **2019**, *7*, 556–559.
82. Amani, M.; Kakooei, M.; Moghimi, A.; Ghorbanian, A.; Ranjgar, B.; Mahdavi, S.; Davidson, A.; Fisette, T.; Rollin, P.; Brisco, B.; et al. Application of google earth engine cloud computing platform, sentinel imagery, and neural networks for crop mapping in Canada. *Remote Sens.* **2020**, *12*, 3561. [CrossRef]
83. Sathiaraj, D.; Huang, X.; Chen, J. Predicting climate types for the Continental United States using unsupervised clustering techniques. *Environmetrics* **2019**, *30*, e2524. [CrossRef]
84. Ip, R.H.; Ang, L.M.; Seng, K.P.; Broster, J.C.; Pratley, J.E. Big data and machine learning for crop protection. *Comput. Electron. Agric.* **2018**, *151*, 376–383. [CrossRef]
85. Saggi, M.K.; Jain, S. Reference evapotranspiration estimation and modeling of the Punjab Northern India using deep learning. *Comput. Electron. Agric.* **2019**, *156*, 387–398. [CrossRef]
86. Reddy, P.C.; Sureshbabu, A. An adaptive model for forecasting seasonal rainfall using predictive analytics. *Int. J. Intell. Eng. Syst.* **2019**, *12*, 22–32. [CrossRef]
87. Sumalatha, M.R.; Akila, M. *Real Time Big Data Analytics for Agricultural Land Hotspot Prediction*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2019; pp. 411–416. [CrossRef]
88. Abbona, F.; Vanneschi, L.; Bona, M.; Giacobini, M. Towards modelling beef cattle management with Genetic Programming. *Livest. Sci.* **2020**, *241*, 104205. .: 10.1016/j.livsci.2020.104205. [CrossRef]
89. Su, Y.; Wang, X. Innovation of agricultural economic management in the process of constructing smart agriculture by big data. *Sustain. Comput. Inform. Syst.* **2021**, *31*, 100579. [CrossRef]
90. Velmurugan, P.; Kannagi, A.; Varsha, M. Superior fuzzy enumeration crop prediction algorithm for big data agriculture applications. *Mater. Today Proc.* **2021**.. [CrossRef]
91. Sitokonstantinou, V.; Drivas, T.; Koukos, A.; Papoutsis, I.; Kontoes, C. Scalable distributed random forest classification for paddy rice mapping. *Zenodo* **2020**, *11*.. [CrossRef]
92. Gumma, M.K.; Thenkabail, P.; Teluguntla, P.; Oliphant, A.; Xiong, J.; Giri, C.; Pyla, V.; Dixit, S.; Whitbread, A. 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 Sens. Taylor Fr.* **2020**, *57*, 302–322. [CrossRef]
93. Mangewa, L.J.; Ndakidemi, P.A.; Alward, R.D.; Kija, H.K.; Bukombe, J.K.; Nasolwa, E.R.; Munishi, L.K. Comparative Assessment of UAV and Sentinel-2 NDVI and GNDVI for Preliminary Diagnosis of Habitat Conditions in Burunge Wildlife Management Area, Tanzania. *Earth* **2022**, *3*, 769–787. [CrossRef]
94. Zhen, Z.; Chen, S.; Yin, T.; Chavanon, E.; Lauret, N.; Guilleux, J.; Henke, M.; Qin, W.; Cao, L.; Li, J.; et al. Using the negative soil adjustment factor of soil adjusted vegetation index (Savi) to resist saturation effects and estimate leaf area index (lai) in dense vegetation areas. *Sensors* **2021**, *21*, 2115. [CrossRef]
95. Aiken, V.C.F.; Dórea, J.R.R.; Acedo, J.S.; de Sousa, F.G.; Dias, F.G.; de Magalhães Rosa, G.J. Record linkage for farm-level data analytics: Comparison of deterministic, stochastic and machine learning methods. *Comput. Electron. Agric.* **2019**, *163*, 104857. . [CrossRef]
96. Amaechi, E.S.; Pham, H.V. 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, Association for Computing Machinery, Haiphong City, Vietnam, 17–19 January 2020; pp. 25–29. [CrossRef]
97. Ye, S.; Liu, D.; Yao, X.; Tang, H.; Xiong, Q.; Zhuo, W.; Song, C. RDCRMG: A raster dataset clean & reconstitution multi-grid architecture for remote sensing monitoring of vegetation dryness. *Remote Sens.* **2018**, *10*, 1376.
98. Xu, C.; Du, X.; Yan, Z.; Fan, X. ScienceEarth: A big data platform for remote sensing data processing. *Remote Sens.* **2020**, *12*, 607. [CrossRef]
99. Sawadogo, P.; Darmont, J. On data lake architectures and metadata management. *J. Intell. Inf. Syst.* **2021**, *56*, 97–120. [CrossRef]
100. Batini, C.; Rula, A.; Scannapieco, M.; Viscusi, G. From data quality to big data quality. *J. Database Manag. JDM* **2015**, *26*, 60–82. [CrossRef]
101. Al-Fuqaha, A.; Guizani, M.; Mohammadi, M.; Aledhari, M.; Ayyash, M. Internet of Things: A survey on enabling technologies, protocols, and applications. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 2347–2376. [CrossRef]
102. Farooq, M.S.; Riaz, S.; Abid, A.; Umer, T.; Zikria, Y. Role of IoT Technology in Agriculture: A Systematic Literature Review. *Electron. Multidiscip. Digit. Publ. Inst.* **2020**, *9*, 319. [CrossRef]