A Selective Survey of Data Mining Techniques in Agriculture

Rudolf Erdei[1,2][0000−0002−3496−3513], Christos Charisis[?][0000−0000−0000−0000], Daniela Delinschi[2][0000−0001−8582−5842], Dimitrios Argyropoulos[?][0000−0000−0000−0000], and Oliviu Matei[1,2][0000−0003−0319−8318]

[1] Technical University of Cluj Napoca, North University Centre of Baia Mare, Romania `oliviu.matei@cunbm.utcluj.ro`

[2] HOLISUN, Romania {`rudolf.erdei,daniela.delinschi`}`@holisun.com`

Abstract. Introduction. This article presents a selective survey on articles for Data Mining in the field of Agriculture. The main purpose of the article is to research what algorithms are the most used in this field for different aspects like temperature, humidity, but also predictions like crop yield and crop health. The research has revealed that certain algorithms, like Support Vector Machines and k-Nearest Neighbour tend to offer better results, so they are favored. Also, in production, ensemble algorithms may offer advantages over using single algorithms. Future research could target distributed learning and other methodologies that improve data safety and privacy, and also improve the cost-effectiveness of Data Discovery Platforms.

1 Introduction

Agriculture, or the domestication of plants, is one of the most important aspects of human activity. With a history of about 15.000 years [1], it surely played a very important role in our history, paving the path to better food and, consequently, for our existence.

Also, agriculture is one of the key fields where technical advances were welcomed and incorporated, leading to improvements in both the quality of life for the workers, as well as the quality of the produce. The invention and improvement of tools, usage of animals, replacement of animals with mechanized machinery, but also intimate knowledge about fertilizers and growth conditions for different plants, all these aspects were refined during the ages, leading to the plentifulness that we have today.

Although famine has not been eradicated, this is not really a production issue, as 30% of the global production is wasted yearly [2], amount that could potentially be used to feed the famine-stricken areas. Hence, the whole global agriculture infrastructure, from the initial seeding and finishing with the dinner plate, can and should be optimized in order to obtain the most out of it. This is even more important if we take into consideration the world population growth [3] and reduction of the land mass that is usable for it [4].

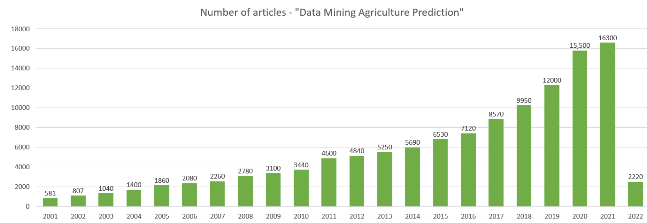

Fig. 1. Number of articles related to data mining used for predictions in agriculture published yearly in 2001 - 2021 (May)

In this context, the Data Revolution [5] that is in full growth today, can and surely will play a major role. The importance of Data Mining and it's usage in agriculture [6] can be assessed also by simply counting the articles that were published in this topic, as can be seen in Figure 1. The trend is currently and exponential one, with no signs of it slowing down over the next decade. Benefits of Data Mining are just beginning to be understood, as algorithms progress and hardware is beginning to be capable enough to process, train and run ever more powerful models with huge amounts of data.

Assessment of the best algorithms for different applications would mean testing these algorithms and selecting the best performing ones. However, this implies a large amount of work, as most of these algorithms do not perform well with the standard parameters, so they would need tuning and optimizations. As part of this work has already been performed, a closer look at the literature will uncover some options that other researchers have taken in order to solve a similar problem (see Table 1. These results could provide as a valuable starting point.

The importance of this topic can also be understood by examining past surveys such as the one by Mucherino et al. [7] in 2009, respectively in 2011 [8], by Raorane et al. in 2013 [9], by Patel et al. in 2014 [10], and by Gandhi et al. in 2016 [11, 12].

Mucherino et al. [13] provide a comprehensive book for the data scientist, that describe Machine Learning (ML) Algorithms as well as Data Mining (DM) techniques that deal with Agriculture. Clustering, Classification and Neural Networks (NN) are all discussed, with specific use cases and ways to implement and optimize the used algorithms.

Cunningham and Holmes [14] discuss different ML algorithms and their usage, using the `WEKA` platform. The described use-case uses a mushroom grading application, with good results. The model's accuracy was comparable to human operators.

Majumdar et al. [15] use the less-known DBSCAN Clustering algorithm to group regions in India with similar rainfall, soil type and temperature ranges, in order to assess which regions have the best wheat crop production. A comprehensive comparison of three clustering methods was done, comparing PAM, CLARA and DBSCAN, with DBSCAN being the best option for this kind of application.

Bauckhage and Kersting [16] discuss about sensor integration, mobility, informational networks in the context of Precision Farming. They present applications and how challenges are addressed when using data mining and pattern recognition in agriculture. This topic is also covered, in less detail, covering only some parts of the ML algorithms, by Yethiraj [17].

Sharma and Mehta [18] describe the architecture and high-level implementation for a Knowledge Management System (KMS), a very useful tool for storing, retrieving and creating new information and knowledge.

Kale and Patil [19] propose a methodology for aiding farmers make informed and correct decisions. Fuzzy logic and machine learning approaches are used to generate expert decisions, thus aiding the farmer and guiding him throughout the farming process.

This article is structured as follows: Section 2 summarizes the architectures used for ML in Agriculture and also depicting a generalization of them. Section 2.1 describes the currently used hardware components for Data Driven Agriculture (DDA), while also exploring possible future research and development trends. Sections 3 to 8 describe the usage of algorithms for different aspects of DDA and provide an overview of them in Section 9 and Table 1. Conclusions and future research are analyzed in Section 10.

2 Architectures of Data Mining Systems used in Agriculture

The architecture of a typical complete data mining process is depicted in figure 2. This class of architectures is very flexible, so it can be completely adapted to the current application, by removing unneeded modules, or by adding others.

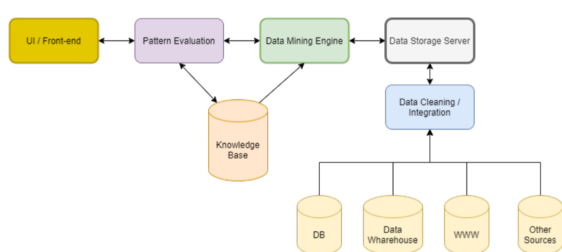

Fig. 2. General Architecture for the Data Mining Process

The IoT context, generally used in agriculture, introduces some constraints for the ML process, primarily caused by the very large amounts of data that one device produces [20]. Multiplying these devices could mean important issues for the cloud platform, that could translate into high costs and/or poor performance.

In this regard, *Federated Learning* [21] and *Collaborative Data Mining* [22] approaches can be used in order to improve both the quality as well as the optimal usage of computing and communication resources.

Concepts like *Digital Twins*, *Edge Computing* (for local data preprocessing), *Context Aware Data Mining* and others, are explored in a humidity forecasting solution for agriculture, by Matei et al. [23]. A large scale agricultural assessment platform is set up, with multiple geographical locations, providing data at a 10 minutes resolution and combining weather forecasts into the assessment, via *Context Aware Data Mining*. The three-tier architecture is scalable and flexible, able to sustain multiple locations, depending on the needs of the overseer [24–27].

2.1 Hardware in Agriculture 4.0

From a hardware point of view, we have many classes of devices that are running inside the Agriculture 4.0 Ecosystem, namely [28]:

1. **Sensors**, that gather data from their environment;
2. **Actuators**, that perform various tasks, like controlling lighting, watering or other parameters;
3. **Edge Computing devices**, that read sensor data, store and prepare it for upload;
4. **Cloud Computing infrastructure**, for performing centralized Data Mining tasks;
5. **User Interface Devices**, like smartphones, laptops, desktop computers, used for interaction with the on-site personnel.

These components and their communication stream are depicted in Figure 3.

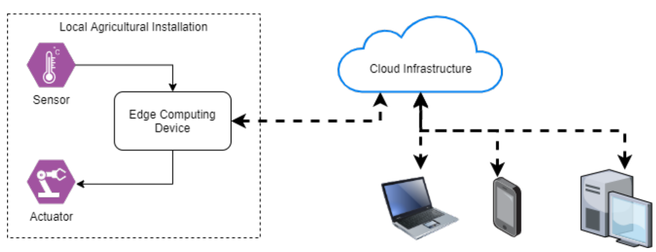

Fig. 3. Hardware components and their connections in Agriculture 4.0

Depending on the scale of the project, not all of these classes of components will be present at any installation. Smaller scale applications require just a few basic components, or merging some of them together (e.g.: ML/DM can be also done on Edge Computing devices, like a Raspberry PI, for a small project).

Gagliardi et al. [29] propose a smart flexible architecture, where drones are used in order to import information from a ZigBee sensor array, that use the

802.15.4 IEEE ZigBee wireless protocol. The Wireless Sensor Network (WSN) [30, 31] does not need to have a direct internet connection, making this solution suitable to extremely remote areas, where access is limited. Specially built drones, with uplink possibilities, are sent at specific intervals (eg.: once a week) in order to connect to the sensor array and upload the data into an SD card. The advantage of this method is obvious: remote locations are easier to manage, as many drones can be programmed to follow a route automatically and there is no line-of-sight needed for manually flying the aircraft. The downside is that the drones introduce a fragile single point of failure, where a significant amount of data can be lost.

Open Source Hardware (OSH) [32–34] is another approach that is suitable for small (or home) projects. Enthusiast builders have a wide range of sensors and components that they can use in order to build a small scale project. OSH can also be used for prototyping, in order to validate certain architectures and software components, due to the low cost and high flexibility that OSH provides. Complete agricultural systems can be easily designed and built, using off-the-shelf components that fit together and are easy to program. 3D printing can also be used in order to build a well-designed case that is cheap and fit for the specific use-case.

Smart Irrigation Systems [35] are easily integrated inside a smart system, due to the increasing availability of energy-efficient and low-cost actuators, in order to optimize both the humidity of the land, providing an optimal environment for the crops to grow, as well as optimize the usage of water, applicable in areas where water is a scarce commodity.

Comprehensive review of hardware usage in smart agriculture was performed by Sharma et al. [36] while some aspects of Smart Agriculture, like protocols, platforms and standards are also discussed by Stočes et al. [37].

In large scale agricultural systems, IoT is beginning to play an important role, by assessing the health of the crops, as well as the growing conditions. By combining this information with weather predictions, more informed choices can be made, while also optimizing the usage of chemicals and irrigation water.

Singh et al. [38] use an approach based on Raspberry PI and Arduino UNO (as the Raspberry PI does not have analog inputs) and an array of off-the-shelf sensors and components to build an automated irrigation system. These systems could be widely used in regions where water is a scarce commodity.

Analyzing the already presented literature regarding hardware usage in agricultural systems, reveals three key aspects of hardware design that must be met by the component, in order to be usable in a long term installation:

- **Power Usage** - advancements in resolution technology for silicon based chips mean better power usage. This also translates in modules, like Bluetooth Low Energy (BLE) Modules, that deliver high-performance at low power usage, enabling data gathering at higher resolutions;
- **Price** - lower prices mean more devices that can be bought with the same budget, enabling higher coverage in the fields, for better control of the input data;

– **Computing power -** higher computing power installed on devices, means more pre-processing can be done at the sensor installation, less information to be uploaded or more functions to be designed, enabling engineers to design better and smarter systems.

Hardware is a very important part of all the platforms for Data Mining in Agriculture. Trends in hardware include the transition to Low-Power Usage, with optimizing both physical components (the circuitry), as well as the software components (algorithmic optimization for low power). Increase in onboard computing power means that more processing will take place at the Edge components, lowering network traffic and costs with cloud infrastructure. Cloud components could potentially be eliminated and replaced with large-scale collaborative distributed systems that are only supervised from a central location.

3 Data mining for soil moisture and air humidity predictions

Soil moisture is an important factor that directly and heavily impacts the quality and performance of crops. Each crop type depends on a specific moisture level in order to perform. While many types of plants can tolerate more drastic moisture changes, some of them are more difficult to grow in less then ideal moisture conditions. In this context, predicting moisture levels enables the farmers to better control growing conditions as well as optimizing water and energy usage.

Matei et al. [39] proposed and built a complete moisture prediction system within the ModSoil Project[3], where a number of 20 weather stations were deployed across the Transylvanian Depression, complete with soil temperature sensors, soil humidity sensor and 10 of them also have rain gauges. Data reading is done with a 10 minutes resolution for each sensor. Preparing the data means computing the average value and the standard deviation for each day. Time Windowing is used, in order to predict the next values, with a Time Window of one day. Several ML algorithms were tested on a smaller scale dataset, including k Nearest Neighbour (k-NN), Support Vector Machines (SVM), Neural Networks (NN), Logistic Regression, Linear Regression, Rule Induction, Fast Large Margin, Decision Tree and Random Forest. The highest accuracy was obtained with k-NN, that yielded a 74.36% accuracy. The average accuracy of the tested algorithms was 68.65% with a standard deviation of 0.033857. In production, the average error was just -0.3%, and the absolute value is 0.68%, proving that k-NN is an algorithm suitable for this application.

Another similar comparison is made by Myers et al. [40] in their work, where they compare a ML approach (without disclosing what algorithms they used), with a satellite based heuristic data processing workflow, the HRLDAS/Noah LSM model. An average 30% improvement was obtained by using the ML approach.

[3] https://research.holisun.com/projects/agriculture-4-0/modsoil

Support Vector Machine (SVM) and Relevance Vector Machine (RVM) algorithms are also used by Hong et al. [41] for predicting soil moisture for `n` days ahead. The developed framework achieves low error rates (15%) and high correlations (95%) when forecasting 14 days ahead.

The scarcity of fresh water is a problem addressed by Singh et al. [38]. This is a huge problem in the context of accelerated desertification [42] that the world is facing, with the global weather change. Singh et al. use a Gradient Boosting Regression Trees (GBRT) algorithm, with very good results, in order to predict the need for an irrigation system to be started in a specific day, so obtaining Water Usage Optimization.

A Deep Learning/Deep Neural Networks (DL/DNN) approach is used by Cai et al. [43] aim to replace empirical formulae and other ML algorithms with the power of Deep Learning, that takes into consideration every available parameter, in order to find hidden links between various input parameters. A Rectified Linear Activation function (ReLU) [44] is used in the neuron, a simple function that improves the speed of data flow inside the Network, improving both the training as well as the exploitation speed. The Adagrad [45] algorithm is used for optimizing the learning stage, an algorithm that actively adjusts the learning rate.

Atanasov [46] predicts the water content of tomatoes by using the leaves color. This method enables the usage of readily available smartphones with high resolution cameras, in order to assess the health and water availability of plants (tomatoes in this case). Seven algorithms are assessed, including ZeroR, Linear Regression, k-NN, M5P Nonlinear Model Tree, Decision Trees, SMOreg and MultyPerc, for deciding the best performing one. The resulting M5P model has a correlation coefficient of 85%.

Aboutalebi et al. [47] present an approach based on multi-spectral imagery, taken with a Unmanned Aerial Vehicle (UAV) to estimate soil moisture at different depths. They assessed multiple algorithms, like Neural Networks and Support Vector Machine (SVM). They found a strong positive correlation of the model accuracy with the site that the training data is sourced, making this approach an ideal candidate for `Federated Learning` or `Collaborative Data Mining`.

Surveys about data mining techniques that use pattern recognition are compiled by Kumar and Kannathasan [48], and also Hemageetha [49]. Both surveys address the problem of soil parameters for agricultural properties, without discussing crop growth or any other crop-related parameters.

A Collaborative Data Mining workflow is also used by Anton et al. [50] for predicting air humidity. This variable directly influences soil moisture, so it's an important parameter that farmers need to take into account. Multiple locations within the Transylvanian Plateau are taken into account in order to successfully raise the accuracy of the resulting predictions. Meanwhile, Muangprathub et al. [51] deploy an array of sensors that also include air humidity, in order to optimize water usage near three villages in India. The enhanced farms are controlled from mobile smartphone applications, enabling low-latency user interaction as well as live system status overview.

3.1 Collaborative Data Mining

`Collaborative Data Mining` (CDM) is a technique meant to improve results obtained from ML models, by using data from other sources, when it is unavailable in a single source. Compared with classical *imputation* of values with medians or deleting entire rows, replacing data with similar data available at other locations has the added benefit of maintaining data resolution as well as variation of the missing data.

Anton et al. [52] use a variety of algorithms, like KNN, LPR, SVM and NN in a collaborative manner to predict soil moisture and temperature, using data from five sources. A comparison, using only KNN, assessing the importance of using multiple data sources, is done in another article by Anton et al. [53], while possible treatments in the case of missing/incomplete data is explored in [54]. If the predictive values obtained in the standalone process are below 0.500 then the collaborative process with multiple sources produces higher values in most cases. If the predictive values obtained in the standalone process are above 0.500 then the collaborative process with multiple sources produces higher values in cases where the predicted source has correlations situated around 0.500.

3.2 Context Aware Data Mining

Data can seldom be regarded as self-sufficient, self-explanatory and free of outside influences. This is the reason why hybridised approaches are used in modern times, in order to improve results and accuracy of models. Matei et al. [55] have proved that infusing contextual data (outside data-sources) into the ML process can improve results. Seemingly inexplicable variations (often regarded as data noise) can otherwise be explained as the result of outside processes or events, that can impact internal data readings. The application that Matei et al. use is based on soil moisture prediction by analysing historical data as well as weather forecast information (linked to temperature and rainfall).

The predicted accuracy of the system is, thus, vastly improved, leading to better use of water resources. A more complete assessment of the results is done by Avram et al. [56]. The importance of context is highlighted by Avram et al. [57], showing that missing contextual information from the dataset has a higher impact in the result than the noise that is present inside the data. The main conclusion of the study is that having a context affected by up to 30% noise and up to 40% missing data, does not influence the results proportionally and the context chosen still proves to be efficient in the noise and missing data ranges that were studied. In other words, a robust context can overcome the shorting of data quantity and quality to a high extent.

3.3 Collaborative vs Context-Aware Data Mining

Rather than being opposite methods, Collaborative Data Mining (CDM) and Context-Aware Data Mining (CADM) are actually complementary in nature. They both improve the quality of resulting models [58], by using missing values from similar sources (CDM) and also infusing contextual (or external) data into the model, like data from public forecasts. Avram et al. [59] propose a practical framework for combining CDM with CADM into a complete knowledge discovery system, with improved accuracy.

3.4 Federated Learning

Two of the largest drawbacks of modern ML systems is the constant need for *data transfer*, *data availability* and, in consequence, the *lack of privacy*. GDPR regulations in Europe have put a strain on Data Driven Systems, making it hard for data consumers to gather user information and behaviour. In this context, *Federated Learning* (FL) approaches might be a viable solution. FL based systems do not centralize data in order to use it, rather build partial models on the edge and then orchestrate the final model build in the cloud without ever transferring private information.

Besides improved data privacy, the Federated Learning approach has the added benefit of distributing the workload of training models to the Edge components of the Cyber-Physical System. Research in the FL field is in its infancy, with some work being done by Zhang et al. [21], with a proposed implementation for *Cyber Physical Systems*, Kumar et al. [60] that present an application of FL for *Smart Agriculture*, and Durrant et al. [61] that use a FL approach for *yield prediction*.

4 Data mining for air temperature

Along with soil moisture, presented in Section 3, *air temperature* is another factor that directly influences growth and well-being of crops. As physics dictates, air temperature also influences soil and air moisture, variation being a key factor into how moisture is transferred to the soil and, finally, to the crops. In this regard, air temperature variation triggers complex agricultural system behaviour, in need of Data Mining techniques to be understood and predicted.

The DBSCAN algorithm is also used by Bilgin et al. [62] in order to pinpoint clusters of similar temperature variations within Turkey. This information can be used to group several locations in order to be able to perform Collaborative Data Mining. The algorithm was used on 66 years of historical data (1930 to 1996). A similar approach was used also by Kohail and El-Halees [63], for the Gaza Strip.

Fathi and Ezziyyani [64] use Unsupervised Learning and Data Mining for predicting air temperature, in order to optimize crop yields, incorporating predictions for the climate changes, for the Morocco region.

A LoRaWAN based system for analytics, including air temperature, is also proposed by Davcev et al. [65]. The system's use case is providing analytics for a grape farm, but, as it also incorporates an important Cloud Computing component, it can be adapted to any other agricultural use-case.

5 Data mining for air CO_2 in agriculture

Unlike animals, that consume O_2 and exhale CO_2, plant life do the opposite, consuming the available CO_2 and producing our much needed O_2. This is why, analyzing the CO_2 level present in the air is important for the well-being of the crops, much more so in the case of closed spaces (vertical agriculture, urban agriculture). Also, CO_2 may be emitted by the soil, raising the importance of green farms and carbon neutral exploitation. Usually, CO_2 alone is not handled by DM methodologies for agriculture, but is used in conjunction with other parameters, presented in this article, in order to build complex DM models, capable to handle and predict more precise data.

Hira and Deshpande [66] present an approach of Data Mining and analysis on Multidimensional Data in agricultural parameters, that also includes CO_2 assessment. Farhate et al. [67] analyze the soil CO_2 emissions for a sugarcane plantation in Brazil, in order to assess its impact on the environment. They assess the relationship between CO_2, soil moisture, soil temperature and others (a total of 18 variables), while also assessing multiple algorithms: Decision Trees, Bayes Classifiers, Neural Networks, Support Vector Machines and Logistic Regression.

Ponce-Guevara et al. [68] present a complete software solution (*GreenFarm-DM*) that integrates Big Data and Data Mining in a green house installation. The system analyzes several factors of crop growth, including soil moisture, temperature and CO_2.

6 Data mining for crop recommendation

Recommendation systems use available knowledge, along with complex assessments of the growing conditions (soil characteristics, weather and climate information) in order to provide the best crop options to the farmers. This assessment targets increases in crop yield and productivity.

Pudulamar et al. [69] use a complete assessment system with the aim of providing accurate crop recommendations. As wrong choice could be costly for the farmers, an ensemble model with majority voting is designed, that uses several algorithms: Random Tree, CHAID, KNN, Naive Bayes. This method raises the accuracy of the prediction to 88%.

Ensemble algorithms are used by Kulkarni et al. [70] and also Akshatha and Shreedhara [71], for crop recommendations for different regions of India. The algorithms used in the ensembles are Random Forest, Naive Bayes, Linear SVM, and K-Nearest Neighbor, individually optimized. This system uses soil types as well as geographic location and local weather data for accurate suggestions of possible cultures. The resulting classification accuracy is 99.91%, a very high value for this kind of problem, mainly due to the ensemble voting mechanism that drastically improves accuracy without compromising flexibility or over-fitting. A similar ensemble system is also developed by Reddy et al. [72], using Random Tree, CHAID, K-Nearest Neighbour (KNN) and Naive Bayes algorithms.

We can observe that, at least for this use-case, ensemble mechanisms provide a more stable system, able to better generalize and better adapt to the data while also maintaining a high accuracy. Ensemble algorithms are, thus, important mechanisms in raising the TRL of the resulting system, making it easier to prepare and deploy the model into production.

An evaluation of several algorithms for optimized crop recommendation is also performed by Arooj et al. [73] for the Pakistan region. They used the WEKA tool and assessed the following algorithms: Decision Trees, Breath First Tree, OneR and Naive Bayes, with the best performing being Naive Bayes.

7 Data mining for crop disease prediction

Crop diseases are a very common problem in the agricultural domain. They can be caused by insects, viruses, bacteria and even other plant life (known as weeds). The impact of these diseases can be dramatic, as entire crops can be lost. In order to prevent this from happening, the farmer needs to know beforehand the complete growing conditions for the selected crop as well as potential problems that may arise.

Ayub and Moqurrab [74] developed an ensemble system using Decision Tree, Random Forest, Neural Networks, Naive Bayes, Support Vector Machines and K-Nearest Neighbor algorithms in order to predict loss due to the grass grub insect. Kumar and Kumar [75] propose another approach, by designing a detection system that uses image processing and data mining in an effort to assess what exactly is threatening the current crop. Their use-case is for the paddy fields in Tamilnadu, Southern India.

Another approach, using an entire *wireless sensor network*, is presented by Tripathy et al. [76]. Their system was trained on historic data and is able to generate almost real-time decisions for disease prediction and mitigation on existing crops, an improvement for the farmers that can result in more optimal usage of time and resources, while also maintaining a greener yield.

8 Data mining in crop yield predictions

A short survey just for crop yield predictions was done by Medar and Rajpurohit in 2014 [77]. An interesting approach to crop yields is proposed by Marinkovic et al. [78], in an application incorporating both Data Mining (using the M5P algorithm) as well as Genetic Algorithms, in an effort to optimize parameters for the best possible yield.

Gupta et al. [79] use Data Mining algorithms such as Random forest, KNN, SVM to monitor and predict some parameters related to yield prediction. Growing parameters like humidity, temperature, soil characteristics are continuously monitored through a sensor network, and prediction results are presented to the farmer via a web interface.

Naive Bayes and K-Nearest Neighbor methods are used by Paul et al. [80] to predict the soil category, in order to estimate crop yields.

Other work include Mishra et al. [81], Guo and Xue [82] and also Dahikar et al. [83] that use Neural Networks in order to compare spatial to temporal approaches. A more extensive review in the use of Neural Networks is done by Khairunniza et al. [84].

9 Data Mining algorithms used in agriculture

A short overview of the ML Algorithms successfully used in agriculture is described in Table 1. This table is useful in the initial stages of researching a particular algorithm for an application, algorithms that offer superior performance.

Abbreviations used in Table 1: *Crop Rec* - Crop Recommendation, *CDP* - Crop Disease Prediction, *SVM* - Support Vector Machines, *SMP* - Soil Moisture Prediction, *CYP* - Crop Yield Prediction, *RVM* - Relevance Vector Machine

As shown in Table 1, a rather large body of work was dedicated to Soil Moisture Prediction, with the most important algorithms being Decision Trees (DT), k-Nearest Neighbour, Bayesian Network and Neural Networks.

10 Conclusions

Agriculture is historically a very important aspect of human existence. It is only natural that a large body of work is dedicated to research in it, incorporated in most of today's research fields (including Electronics, Computer Science and Engineering). Today's Computer Science research landscape for Agriculture proposes a vast number of mature algorithms that can be used to develop high TRL production systems.

In Data Driven Agriculture, predictions have been made (or proven possible), for: soil and air humidity, soil and air temperature, CO_2/O_2, crop disease prediction, crop recommendation and even crop yield prediction. These aspects are highly important, in the context of optimizing the outcome of agricultural activities (maximizing outcome, minimizing waste and pollution due to chemical treatments). Monitoring and modifying these aspects leads to less usage of said chemicals, so it can also define a crop as being *BIO*, in the context of raised awareness and interest on this topic.

10.1 Future Research Trends

Researching the available literature, it can be noticed that some algorithms stand out, like *SVM*, *kNN* and *Decision Trees*. These are highly used and could form a good basis for applicative research on a high TRL ensemble platform that would be suitable for a production-grade Data Mining Platform. Ensemble algorithms provide better model accuracy, needed for a production system.

On the topic of hardware, further research is needed for advanced algorithms for dimensionality reduction of the gathered data, in order to better be able

Table 1. Overview of Data Mining Algorithms used in Agriculture

Algorithm	Used by	Applications
SVM	[39–41, 47, 66, 67, 70, 71] [74, 79, 85–91]	SMP, CYP,Air CO_2,CDP
RVM	[41]	SMP
Deep Learning (DL)	[43]	SMP
Naive Bayes	[69–74, 80]	Crop Rec , CDP, CYP
K-means Clustering	[90]	SMP
k-Nearest Neighbour	[39, 46, 69–72, 74, 79, 80] [90–95]	SMP, CYP, CDP
Neural Networks	[39, 47, 66, 67, 74, 81–84]	SMP, Air CO_2, CDP, CYP
Logistic Regression	[39, 66, 67]	SMP, Air CO_2
Linear Regression	[39, 46]	SMP
ZeroR, SMOreg, MultyPerc	[46]	SMP
Rule Induction	[39]	SMP
Fast Large Margin	[39]	SMP
Random Forest	[39, 69–72, 74, 79]	SMP, Crop Rec, CDP, CYP
Decision Tree Analysis	[46, 66, 67, 73, 74, 78, 91] [93]	SMP, CYP, Air CO_2, CDP
GBRT	[38]	SMP
Bayesian network	[66, 67, 90]	SMP, Crop Rec, Air CO_2
Unsupervised Clustering	[64]	Air Temp
DBSCAN	[62, 63]	Air Temp
LoRaWAN	[65]	Air Temp
CHAID	[69, 72]	Crop Rec
OneR	[73]	Crop Rec
DM Survey	[48, 49]	-

to maintain its intrinsic information, while lowering as much as possible the required bandwidth and storage requirements.

Today's trends include many Distributed Learning approaches and methodologies, many of them being in initial phases. More work is to be done especially on Federated Learning, but also the possibility of integration with other approaches, like Collaborative and Context-Aware Data Mining.

References

1. Daniel E Vasey. *An ecological history of agriculture 10,000 BC-AD 10,000*. Purdue University Press, 2002.
2. Meera Verma. Food wastage—energy wasted. In *Energy use in global food production*, pages 23–26. Springer, 2015.
3. Max Roser. Future population growth. *Our world in data*, 2013.
4. Prue Campbell. The future prospects for global arable land. *Global Food and Water Crises Research Programme. Strategic analysis paper*, pages 2–6, 2011.
5. Liran Einav and Jonathan Levin. The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1):1–24, 2014.
6. B Milovic, V Radojevic, et al. Application of data mining in agriculture. *Bulgarian Journal of Agricultural Science*, 21(1):26–34, 2015.
7. Antonio Mucherino, Petraq Papajorgji, and Panos M Pardalos. A survey of data mining techniques applied to agriculture. *Operational Research*, 9(2):121–140, 2009.
8. Antonio Mucherino and Georg Ruß. Recent developments in data mining and agriculture. In *Industrial Conference on Data Mining-Workshops*, pages 90–98, 2011.
9. AA Raorane and RV Kulkarni. Role of data mining in agriculture. *International Journal of Computer Scienceand Information Technologies*, 4(2):270–272, 2013.
10. Hetal Patel and Dharmendra Patel. A brief survey of data mining techniques applied to agricultural data. *International Journal of Computer Applications*, 95(9), 2014.
11. Niketa Gandhi and Leisa J Armstrong. A review of the application of data mining techniques for decision making in agriculture. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 1–6. IEEE, 2016.
12. D Sabareeswaran and A Edwin Robert. A survey on data mining techniques in agriculture. *Indian J. Innov. Dev.*, 5:8, 2016.
13. Antonio Mucherino, Petraq Papajorgji, and Panos M Pardalos. *Data mining in agriculture*, volume 34. Springer Science & Business Media, 2009.
14. Sally Jo Cunningham and Geoffrey Holmes. Developing innovative applications in agriculture using data mining. In *The proceedings of the Southeast Asia regional computer confederation conference*, pages 25–29. Citeseer, 1999.
15. Jharna Majumdar, Sneha Naraseeyappa, and Shilpa Ankalaki. Analysis of agriculture data using data mining techniques: application of big data. *Journal of Big data*, 4(1):1–15, 2017.
16. Christian Bauckhage and Kristian Kersting. Data mining and pattern recognition in agriculture. *KI-Künstliche Intelligenz*, 27(4):313–324, 2013.
17. NG Yethiraj. Applying data mining techniques in the field of agriculture and allied sciences. *International Journal of Business Intelligents ISSN*, pages 2278–2400, 2012.

18. Latika Sharma and Nitu Mehta. Data mining techniques: A tool for knowledge management system in agriculture. *International Journal of scientific and technology research*, 1(5):67–73, 2012.
19. Shivani S Kale and Preeti S Patil. Data mining technology with fuzzy logic, neural networks and machine learning for agriculture. In *Data management, analytics and innovation*, pages 79–87. Springer, 2019.
20. Zhaoyu Zhai, José Fernán Martínez, Victoria Beltran, and Néstor Lucas Martínez. Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170:105256, 2020.
21. Chong Zhang, Xiao Liu, Xi Zheng, Rui Li, and Huai Liu. Fenghuolun: A federated learning based edge computing platform for cyber-physical systems. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 1–4. IEEE, 2020.
22. Oliviu Matei, Carmen Anton, Sebastian Scholze, and Claudio Cenedese. Multi-layered data mining architecture in the context of internet of things. In *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, pages 1193–1198. IEEE, 2017.
23. O Matei, C Anton, A Bozga, and P Pop. Multi-layered architecture for soil moisture prediction in agriculture 4.0. In *Proceedings of international conference on computers and industrial engineering, CIE*, volume 2, pages 39–48, 2017.
24. Raimo Nikkilä, Ilkka Seilonen, and Kari Koskinen. Software architecture for farm management information systems in precision agriculture. *Computers and electronics in agriculture*, 70(2):328–336, 2010.
25. JA López-Riquelme, N Pavón-Pulido, H Navarro-Hellín, F Soto-Valles, and R Torres-Sánchez. A software architecture based on fiware cloud for precision agriculture. *Agricultural water management*, 183:123–135, 2017.
26. Li Tan, Ronald Haley, Riley Wortman, and Qin Zhang. An extensible and integrated software architecture for data analysis and visualization in precision agriculture. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, pages 271–278. IEEE, 2012.
27. Qing Zhu, Kaihua Liao, Yan Xu, Guishan Yang, Shaohua Wu, and Shenglu Zhou. Monitoring and prediction of soil moisture spatial–temporal variations from a hydropedological perspective: a review. *Soil Research*, 50(8):625–637, 2013.
28. Parijata Majumdar, Sanjoy Mitra, and Diptendu Bhattacharya. Iot for promoting agriculture 4.0: a review from the perspective of weather monitoring, yield prediction, security of wsn protocols, and hardware cost analysis. *Journal of Biosystems Engineering*, pages 1–22, 2021.
29. G. Gagliardi, M. Lupia, G. Cario, F. C. Gaccio, V. D'angelo, A. I. M. Cosma, and A. Casavola. An internet of things solution for smart agriculture. *Agronomy*, 11(11), 2021.
30. Mohamed Rawidean Mohd Kassim, Ibrahim Mat, and Ahmad Nizar Harun. Wireless sensor network in precision agriculture application. In *2014 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–5. IEEE, 2014.
31. Denis Ilie-Ablachim, George Cristian Pătru, Iulia-Maria Florea, and Daniel Rosner. Monitoring device for culture substrate growth parameters for precision agriculture: Acronym: Monisen. In *2016 15th RoEduNet Conference: Networking in Education and Research*, pages 1–7. IEEE, 2016.
32. FJ Mesas-Carrascosa, D Verdú Santano, JE Meroño, M Sánchez De La Orden, and A García-Ferrer. Open source hardware to monitor environmental parameters in precision agriculture. *Biosystems engineering*, 137:73–83, 2015.

33. Spyros Fountas, Søren Marcus Pedersen, and Simon Blackmore. Ict in precision agriculture–diffusion of technology. *ICT in agriculture: perspective of technological innovation, E. Gelb and A. Offer (eds), http://departments. agri. huji. ac. il/economics/gelb-main. html*, 2005.
34. Brett Whelan and James Taylor. *Precision agriculture for grain production systems*. Csiro publishing, 2013.
35. Sanbo Li. Application of the internet of things technology in precision agriculture irrigation systems. In *2012 International Conference on Computer Science and Service System*, pages 1009–1013. IEEE, 2012.
36. Brij Bhushan Sharma and Nagesh Kumar. Internet of things-based hardware and software for smart agriculture: A review. *Proceedings of ICRIC 2019*, pages 151–157, 2020.
37. Michal Stoces, Jiri Vanek, Jan Masner, and Jan Pavlik. Internet of things (iot) in agriculture-selected aspects. *Agris on-line Papers in Economics and Informatics*, 8(665-2016-45107):83–88, 2016.
38. Gursimran Singh, Deepak Sharma, Amarendra Goap, Sugandha Sehgal, AK Shukla, and Satish Kumar. Machine learning based soil moisture prediction for internet of things based smart irrigation system. In *2019 5th International Conference on Signal Processing, Computing and Control (ISPCC)*, pages 175–180. IEEE, 2019.
39. Oliviu Matei, Teodor Rusu, Adrian Petrovan, and Gabriel Mihuţ. A data mining system for real time soil moisture prediction. *Procedia Engineering*, 181:837–844, 2017.
40. William Myers, Seth Linden, and Gerry Wiener. A data mining approach to soil temperature and moisture prediction. *National Center for Atmospheric Research, Boulder, CO*, 2009.
41. Zhihao Hong, Z Kalbarczyk, and Ravishankar K Iyer. A data-driven approach to soil moisture collection and prediction. In *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 1–6. IEEE, 2016.
42. Eleonora Grilli, Sílvia CP Carvalho, Tommaso Chiti, Elio Coppola, Rosaria D'Ascoli, Tommaso La Mantia, Rossana Marzaioli, Micòl Mastrocicco, Fernando Pulido, Flora Angela Rutigliano, et al. Critical range of soil organic carbon in southern europe lands under desertification risk. *Journal of Environmental Management*, 287:112285, 2021.
43. Yu Cai, Wengang Zheng, Xin Zhang, Lili Zhangzhong, and Xuzhang Xue. Research on soil moisture prediction model based on deep learning. *PloS one*, 14(4):e0214508, 2019.
44. Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.
45. Agnes Lydia and Sagayaraj Francis. Adagrad—an optimizer for stochastic gradient descent. *Int. J. Inf. Comput. Sci*, 6(5), 2019.
46. SS Atanasov. Predicting soil moisture based on the color of the leaves using data mining and machine learning techniques. In *IOP Conference Series: Materials Science and Engineering*, volume 1031, page 012076. IOP Publishing, 2021.
47. Mahyar Aboutalebi, L Niel Allen, Alfonso F Torres-Rua, Mac McKee, and Calvin Coopmans. Estimation of soil moisture at different soil levels using machine learning techniques and unmanned aerial vehicle (uav) multispectral imagery. In *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping IV*, volume 11008, page 110080S. International Society for Optics and Photonics, 2019.

48. A Kumar and N Kannathasan. A survey on data mining and pattern recognition techniques for soil data mining. *IJCSI International Journal of Computer Science Issues*, 8(3), 2011.
49. N Hemageetha. A survey on application of data mining techniques to analyze the soil for agricultural purpose. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 3112–3117. IEEE, 2016.
50. Carmen Ana Anton, Anca Avram, Adrian Petrovan, and Oliviu Matei. Performance analysis of collaborative data mining vs context aware data mining in a practical scenario for predicting air humidity. In *Proceedings of the Computational Methods in Systems and Software*, pages 31–40. Springer, 2019.
51. Jirapond Muangprathub, Nathaphon Boonnam, Siriwan Kajornkasirat, Narongsak Lekbangpong, Apirat Wanichsombat, and Pichetwut Nillaor. Iot and agriculture data analysis for smart farm. *Computers and electronics in agriculture*, 156:467–474, 2019.
52. Carmen Ana Anton, Oliviu Matei, and Anca Avram. Collaborative data mining in agriculture for prediction of soil moisture and temperature. In *Computer Science On-Line Conference*, pages 141–151. Springer, 2019.
53. Carmen Anton, Oliviu Matei, and Anca Avram. Use of multiple data sources in collaborative data mining. In *Proceedings of the Computational Methods in Systems and Software*, pages 189–198. Springer, 2019.
54. Carmen Ana Anton, Oliviu Matei, and Anca Avram. Missing data in collaborative data mining. In *Proceedings of the Computational Methods in Systems and Software*, pages 100–109. Springer, 2019.
55. Oliviu Matei, Teodor Rusu, Andrei Bozga, Petrica Pop-Sitar, and Carmen Anton. Context-aware data mining: embedding external data sources in a machine learning process. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 415–426. Springer, 2017.
56. Anca Avram, Oliviu Matei, Camelia-M Pintea, and Petrica C Pop. Context quality impact in context-aware data mining for predicting soil moisture. *Cybernetics and Systems*, 51(7):668–684, 2020.
57. Anca Avram, Oliviu Matei, Camelia-M Pintea, Petrica C Pop, and Carmen Ana Anton. How noisy and missing context influences predictions in a practical context-aware data mining system. In *International Workshop on Soft Computing Models in Industrial and Environmental Applications*, pages 22–32. Springer, 2020.
58. Anca Avram, Oliviu Matei, Camelia-M Pintea, Petrica C Pop, and Carmen Ana Anton. Context-aware data mining vs classical data mining: case study on predicting soil moisture. In *International Workshop on Soft Computing Models in Industrial and Environmental Applications*, pages 199–208. Springer, 2019.
59. Anca Avram, Oliviu Matei, Camelia Pintea, and Carmen Anton. Innovative platform for designing hybrid collaborative & context-aware data mining scenarios. *Mathematics*, 8(5):684, 2020.
60. Prabhat Kumar, Govind P Gupta, and Rakesh Tripathi. Pefl: Deep privacy-encoding based federated learning framework for smart agriculture. *IEEE Micro*, 2021.
61. Aiden Durrant, Milan Markovic, David Matthews, David May, Jessica Enright, and Georgios Leontidis. The role of cross-silo federated learning in facilitating data sharing in the agri-food sector. *Computers and Electronics in Agriculture*, 193:106648, 2022.
62. T Tugay Bilgin and A Yılmaz Çamurcu. A data mining application on air temperature database. In *International Conference on Advances in Information Systems*, pages 68–76. Springer, 2004.

63. Sarah N Kohail and Alaa M El-Halees. Implementation of data mining techniques for meteorological data analysis. *Intl. Journal of Information and Communication Technology Research (JICT)*, 1(3), 2011.
64. Maroi Tsouli Fathi and Mostafa Ezziyyani. How can data mining help us to predict the influence of climate change on mediterranean agriculture? *International Journal of Sustainable Agricultural Management and Informatics*, 5(2-3):168–180, 2019.
65. Danco Davcev, Kosta Mitreski, Stefan Trajkovic, Viktor Nikolovski, and Nikola Koteli. Iot agriculture system based on lorawan. In *2018 14th IEEE International Workshop on Factory Communication Systems (WFCS)*, pages 1–4. IEEE, 2018.
66. Swati Hira and PS Deshpande. Data analysis using multidimensional modeling, statistical analysis and data mining on agriculture parameters. *Procedia Computer Science*, 54:431–439, 2015.
67. Camila Viana Vieira Farhate, Zigomar Menezes de Souza, Stanley Robson de Medeiros Oliveira, Rose Luiza Moraes Tavares, and João Luís Nunes Carvalho. Use of data mining techniques to classify soil co2 emission induced by crop management in sugarcane field. *Plos one*, 13(3):e0193537, 2018.
68. KL Ponce-Guevara, JA Palacios-Echeverría, E Maya-Olalla, HM Domínguez-Limaico, LE Suárez-Zambrano, PD Rosero-Montalvo, DH Peluffo-Ordóñez, and JC Alvarado-Pérez. Greenfarm-dm: A tool for analyzing vegetable crops data from a greenhouse using data mining techniques (first trial). In *2017 IEEE second ecuador technical chapters meeting (ETCM)*, pages 1–6. IEEE, 2017.
69. S Pudumalar, E Ramanujam, R Harine Rajashree, C Kavya, T Kiruthika, and J Nisha. Crop recommendation system for precision agriculture. In *2016 Eighth International Conference on Advanced Computing (ICoAC)*, pages 32–36. IEEE, 2017.
70. Nidhi H Kulkarni, GN Srinivasan, BM Sagar, and NK Cauvery. Improving crop productivity through a crop recommendation system using ensembling technique. In *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, pages 114–119. IEEE, 2018.
71. KR Akshatha and KS Shreedhara. Implementation of machine learning algorithms for crop recommendation using precision agriculture. *International Journal of Research in Engineering, Science and Management (IJRESM)*, 1(6):58–60, 2018.
72. D Anantha Reddy, Bhagyashri Dadore, and Aarti Watekar. Crop recommendation system to maximize crop yield in ramtek region using machine learning. *International Journal of Scientific Research in Science and Technology*, 6(1):485–489, 2019.
73. Ansif Arooj, Mohsin Riaz, and Malik Naeem Akram. Evaluation of predictive data mining algorithms in soil data classification for optimized crop recommendation. In *2018 International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–6. IEEE, 2018.
74. Umair Ayub and Syed Atif Moqurrab. Predicting crop diseases using data mining approaches: classification. In *2018 1st International Conference On Power, Energy And Smart Grid (Icpesg)*, pages 1–6. IEEE, 2018.
75. SA Ramesh Kumar and K Ramesh Kumar. A study on paddy crops disease prediction using data mining techniques. *crops*, 1:2, 2015.
76. AK Tripathy, J Adinarayana, D Sudharsan, SN Merchant, UB Desai, K Vijayalakshmi, D Raji Reddy, G Sreenivas, S Ninomiya, M Hirafuji, et al. Data mining and wireless sensor network for agriculture pest/disease predictions. In *2011 World Congress on Information and Communication Technologies*, pages 1229–1234. IEEE, 2011.

77. Ramesh A Medar and Vijay S Rajpurohit. A survey on data mining techniques for crop yield prediction. *International Journal of Advance Research in Computer Science and Management Studies*, 2(9):59–64, 2014.
78. Branko Marinković, Jovan Crnobarac, Sanja Brdar, Borislav Antić, Goran Jaćimović, and Vladimir Crnojević. Data mining approach for predictive modeling of agricultural yield data. In *Proc. First Int Workshop on Sensing Technologies in Agriculture, Forestry and Environment (BioSense09), Novi Sad, Serbia*, pages 1–5, 2009.
79. Gautam Gupta, Rashi Setia, Archana Meena, and Bhavnesh Jaint. Environment monitoring system for agricultural application using iot and predicting crop yield using various data mining techniques. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 1019–1025. IEEE, 2020.
80. Monali Paul, Santosh K Vishwakarma, and Ashok Verma. Analysis of soil behaviour and prediction of crop yield using data mining approach. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 766–771. IEEE, 2015.
81. Shruti Mishra, Priyanka Paygude, Snehal Chaudhary, and Sonali Idate. Use of data mining in crop yield prediction. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pages 796–802. IEEE, 2018.
82. William W Guo and Heru Xue. Crop yield forecasting using artificial neural networks: A comparison between spatial and temporal models. *Mathematical Problems in Engineering*, 2014, 2014.
83. Snehal S Dahikar and Sandeep V Rode. Agricultural crop yield prediction using artificial neural network approach. *International journal of innovative research in electrical, electronics, instrumentation and control engineering*, 2(1):683–686, 2014.
84. Siti Khairunniza-Bejo, Samihah Mustaffha, and Wan Ishak Wan Ismail. Application of artificial neural network in predicting crop yield: A review. *Journal of Food Science and Engineering*, 4(1):1, 2014.
85. M Kashif Gill, Tirusew Asefa, Mariush W Kemblowski, and Mac McKee. Soil moisture prediction using support vector machines 1. *JAWRA Journal of the American Water Resources Association*, 42(4):1033–1046, 2006.
86. Qian Zhu, Yushi Wang, and Yulin Luo. Improvement of multi-layer soil moisture prediction using support vector machines and ensemble kalman filter coupled with remote sensing soil moisture datasets over an agriculture dominant basin in china. *Hydrological Processes*, 35(4):e14154, 2021.
87. M Kashif Gill, Mariush W Kemblowski, and Mac McKee. Soil moisture data assimilation using support vector machines and ensemble kalman filter 1. *JAWRA Journal of the American Water Resources Association*, 43(4):1004–1015, 2007.
88. Bushra Zaman, Mac McKee, and Christopher MU Neale. Fusion of remotely sensed data for soil moisture estimation using relevance vector and support vector machines. *International journal of remote sensing*, 33(20):6516–6552, 2012.
89. Di Liu, Ashok K Mishra, and Zhongbo Yu. Evaluating uncertainties in multilayer soil moisture estimation with support vector machines and ensemble kalman filtering. *Journal of Hydrology*, 538:243–255, 2016.
90. CN Vanitha, N Archana, and R Sowmiya. Agriculture analysis using data mining and machine learning techniques. In *2019 5th international conference on advanced computing & communication systems (ICACCS)*, pages 984–990. IEEE, 2019.
91. Alberto Gonzalez-Sanchez, Juan Frausto-Solis, and Waldo Ojeda-Bustamante. Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, 12(2):313–328, 2014.

92. Sevim Seda Yamaç, Cevdet Şeker, and Hamza Negiş. Evaluation of machine learning methods to predict soil moisture constants with different combinations of soil input data for calcareous soils in a semi arid area. *Agricultural Water Management*, 234:106121, 2020.
93. Odi Nurdiawan, DA Kurnia, Dodi Solihudin, Tuti Hartati, and Tati Suprapti. Comparison of the k-nearest neighbor algorithm and the decision tree on moisture classification. In *IOP Conference Series: Materials Science and Engineering*, volume 1088, page 012031. IOP Publishing, 2021.
94. Xiangyan Meng, Zhongxue Zhang, and Xinying Xu. A novel k-nearest neighbor algorithm based on i-divergence with application to soil moisture estimation in maize field. *J. Softw.*, 9(4):841–846, 2014.
95. P Salma Khatoon and Muqeem Ahmed. A crop selection framework using k nearest neighbour for iot semantic interoperability applications. In *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 262–267. IEEE, 2021.