International Journal of Gender, Technology
..ISSN: 2040-0748

UGC Care Journal
Vol-12 Issue-02 Dec 2023

Crop yield prediction using machine learning: A systematic literature review

[1]LIKITHASAI BOINA, [2]DINESH KUMAR, [3]B AJAY CHALLA, [4]S PAVAN,
[5]Mrs. V. Mounika (Assistant Professor),
U.G.Students,
Department of CSE,
MALLA REDDY INSTITUTE OF TECHNOLOGY & SCIENCE, Maisammaguda, Medchal (M), Hyderabad-500100, T. S.

ABSTRACT

When making choices about which crops to produce and how to care for them throughout their growth, machine learning is a crucial decision-support tool for predicting crop yields. Researchers have used a number of machine learning algorithms to help in agricultural production prediction. We extracted and synthesized the techniques and characteristics utilized in agricultural yield prediction research by conducting a Systematic Literature Review (SLR). We used inclusion and exclusion criteria to narrow the search results down to 50 papers from 567 that met our criteria across six internet databases. We provide detailed analyses of the characteristics and methodologies employed in these chosen studies and recommendations for further study. Our data shows that temperature, rainfall, and soil type are the most often utilized features in these models, with Artificial Neural Networks being the most often employed technique. We then conducted a second search in electronic databases to find research based on deep learning, reached 30 publications, and extracted the used deep learning algorithms. This finding was based on the examination of 50 papers based on machine learning. In these research, Convolutional Neural Networks (CNN) were the most often utilized deep learning method, followed by Deep Neural Networks (DNN) and Long Short Term Memory (LSTM).

1. Introduction

Machine learning (ML) approaches are used in many fields, ranging from supermarkets to evaluate the behavior of customers (Ayodele, 2010) to the prediction of customers' phone use (Witten et al., 2016). Machine learning is also being used in agriculture for several years (McQueen et al., 1995). Crop yield prediction is one of the challenging problems in precision agriculture, and many models have been proposed and validated so far. This problem requires the use of several datasets since crop yield depends on many different factors such as climate, weather, soil, use of fertilizer, and seed variety (Xu et al., 2019). This indicates that crop yield prediction is not a trivial task; instead, it consists of several complicated steps. Nowadays, crop yield prediction models can estimate the actual yield reasonably, but a better performance in yield prediction is still desirable (Filippi et al., 2019a). Machine learning, which is a branch of Artificial Intelligence (AI) focusing on learning, is a practical approach that can provide better yield prediction based on several features. Machine learning (ML) can determine patterns and correlations and discover knowledge from datasets.

The models need to be trained using datasets, where the outcomes are represented based on past experience. The predictive model is built using several features, and as such, parameters of the models are determined using historical data during the training phase. For the testing phase, part of the historical data that has not been used for training is used for the performance evaluation purpose. An ML model can be descriptive or predictive, depending on the research

problem and research questions. While descriptive models are used to gain knowledge from the collected data and explain what has happened, predictive models are used to make predictions in the future (Alpaydin, 2010). ML studies consist of different challenges when aiming to build a high-performance predictive model. It is crucial to select the right algorithms to solve the problem at hand, and in addition, the algorithms and the underlying platforms need to be capable of handling the volume of data. To get an overview of what has been done on the application of ML in crop yield prediction, we performed a systematic literature review (SLR). A Systematic Literature Review (SLR) shows the potential gaps in research on a particular area of problem and guides both practitioners and researchers who wish to do a new research study on that problem area. By following a methodology in SLR, all relevant studies are accessed from electronic databases, synthesized, and presented to respond to research questions defined in the study. An SLR study leads to new perspectives and helps new researchers in the field to understand the state-of-the-art.

All procedures must be properly described and the outcomes must be open and accessible to other researchers in order for an SLR study to be considered repeatable. Objectivity and transparency are key components of a well-conducted SLR research (Kitchenham et al., 2007). An SLR, as the name suggests, must be comprehensive and systematically review all available material. To far, there has been no comprehensive review of the literature on the topic of using machine learning to forecast agricultural yields; however, that has all changed now. Our study's empirical findings and answers to the research questions posed in this review article are presented below. Here is how the rest of the paper is structured: Section 2 provides an overview of the context. The approach is covered in Section 3. The findings of the SLR are detailed in Section 4. Section 5 delves into the study on agricultural production prediction using deep learning. Section 5 showcases the conversation, and Section 7 brings this paper to a close.

2. Related work

Crop yield prediction is an essential task for the decision-makers at national and regional levels (e.g., the EU level) for rapid decisionmaking. An accurate crop yield prediction model can help farmers to decide on what to grow and when to grow. There are different approaches to crop yield prediction. This review article has investigated what has been done on the use of machine learning in crop yield prediction in the literature. During our analysis of the retrieved publications, one of the exclusion criteria is that the publication is a survey or traditional review paper. Those excluded publications are, in fact, related work and are discussed in this section. Chlingaryan and Sukkarieh performed a review study on nitrogen status estimation using machine learning (Chlingaryan et al., 2018). The paper concludes that quick developments in sensing technologies and ML techniques will result in cost-effective solutions in the agricultural sector. Elavarasan et al. performed a survey of publications on machine learning models associated with crop yield prediction based on climatic parameters. The paper advises looking broad to find more parameters that account for crop yield (Elavarasan et al., 2018). Liakos et al. (2018) published a review paper on the application of machine learning in the agricultural sector. The analysis was performed with publications focusing on crop management, livestock management, water management, and soil management. Li, Lecourt, and Bishop performed a review study on determining the ripeness of fruits to decide the optimal harvest time and yield prediction (Li et al., 2018). Mayuri and Priya addressed the challenges and methodologies that are encountered in the field of image processing and machine learning in the agricultural sector and especially in the detection of diseases (Mayuri and Priya, xxxx). Somvanshi and Mishra presented several machine learning approaches and their application in plant biology (Somvanshi and Mishra, 2015).

Gandhi and Armstrong published a review paper on the application of data mining in the agricultural sector in general, dealing with decision making. They concluded that further research needs to be done to see how the implementation of data mining into complex agricultural datasets could be realized (Gandhi and Armstrong, 2016). Beulah performed a survey on the various data mining techniques that are used for crop yield prediction and concluded that the crop yield prediction could be solved by employing data mining

techniques (Beulah, 2019). According to our survey of review articles, the significant ones of which are presented in this section, this paper is the first SLR that focuses on the application of machine learning in the crop yield prediction problem. The existing survey studies did not systematically review the literature, and most of them reviewed studies on a specific aspect of crop yield prediction. Also, we presented 30 deep learning-based studies in this article and discussed which deep learning algorithms have been used in these studies.

3. Methodology

3.1. Review protocol

Before conducting the systematic review, a review protocol is defined. The review has been done using the well-known review guidelines provided by Kitchenham et al. (2007). Firstly, the research questions are defined. When research questions are ready, databases are used to select the relevant studies. The databases that were used in this study are Science Direct, Scopus, Web of Science, Springer Link, Wiley, and Google Scholar. After the selection of relevant studies, they were filtered and assessed using a set of exclusion and quality criteria. All the relevant data from the selected studies are extracted, and eventually, the extracted data were synthesized in response to the research questions. The approach we followed can be split up into three parts: plan review, conduct review, and report review. The first stage is planning the review. In this stage, research questions are identified, a protocol is developed, and eventually, the protocol is validated to see if the approach is feasible. In addition to the research questions, publication venues, initial search strings, and publication selection criteria are also defined. When all of this information is defined, the protocol is revised one more time to see if it represents a proper review protocol. In Fig. 1, the internal steps of the Plan Review stage are represented. The second stage is conducting the review, which is represented in Fig. 2. When conducting the review, the publications were selected by going through all the databases. The data was extracted, which means that their information regarding authors, year of publication, type of publication, and more information regarding the research questions were stored. After all the necessary data was extracted correctly, the data was synthesized in order to provide an overview of the relevant papers published so far. In the final stage, a.k.a., Reporting the Review, the review was concluded by documenting the results and addressing the research questions, as shown in Fig. 3.

domains of application, which implies that there are likely many published research that do not fall within the purview of this review piece. An automated search handles the fundamental searching. "Machine learning" and "yield prediction" were the first search parameters. We looked for alternative terms for the keywords by retrieving relevant articles and reading their abstracts. There were six databases that were searched. Use the search terms "machine learning" and "yield prediction" to get a comprehensive overview of the research. In order to ensure that no relevant studies were missed, a more intricate search string was constructed after the exclusion criteria were applied and all the results were analyzed.

Here is the final search string: (("machine learning" OR "artificial intelligence") AND "data mining" AND ("yield prediction" OR "yield forecasting" OR "yield es timation")). There were 567 studies that were retrieved when this search query was executed. Detailed descriptions of the search strings for each database are given below: Direct from the scientific community: We are looking for results that include both "machine learning" and "yield prediction." (Title, abstract, keywords, and [("machine learning" OR "artificial intelligence") AND "data mining" AND ("yield prediction" OR "yield forecasting" OR "yield estimation")))))](Title, abstract, and important terms)! Index Scopus: Please enter ["machine learning" AND "yield pre diction"] as the search keyword.(Title, abstract, keywords, and [("machine learning" OR "artificial intelligence") AND "data mining" AND ("yield prediction" OR "yield forecasting" OR "yield estimation")))))] (Title, abstract, and important terms)! Web of Science: ["machine learning" AND "yield prediction"] (title, abstract, author keywords, and Keywords Plus) is the search phrase required. Publisher: Springer Enter the following search terms: ["machine learning" AND "yield prediction"](anywhere) and [(("machine learning" OR "artificial intelligence") AND "data mining" AND

("yield prediction" OR "yield forecasting" OR "yield estimation"))](anywhere) Wiley: Machine learning and yield prediction are the search terms (somewhere). The Google Scholar website: A search phrase that may be used anywhere would be ["machine learning" AND "yield prediction"] (anywhere) or [(("machine learning" OR "artificial intelligence") AND "data mining" AND ("yield prediction" OR "yield forecasting" OR "yield estimation"))] (anywhere). No papers were found for the following search strings on Web of Science and Wiley: [(("machine learning" OR "artificial intelligence") AND "data mining" AND ("yield prediction" OR "yield forecasting" OR "yield estimation"))].

Table 1 Distribution of papers based on the databases

Database	# of initially retrieved papers	# of papers after exclusion criteria	Percentage of Papers (%)
Science Direct	17	4	8
Scopus	68	11	22
Web of Science	32	0	0
Springer Link	132	10	20
Wiley	20	1	2
Google Scholar	298	24	48
Total	567	50	100

Criteria for exclusion 2 - The publication does not adhere to the English language standard 3 - Publication that has been previously published or is a duplicate Criteria for exclusion 4. The whole text of the article is unavailable. Criteria for exclusion 5. A review or survey article is published. Criteria for exclusion 6. Publication was released before to 2008. Only 77 papers were retained for further analysis after the application of the first three exclusion criteria. Only fifty research made it through the rigorous screening process that followed the six exclusion criteria. The number of articles that were originally retrieved and the number of papers that were retained after the application of selection criteria are shown in Table 1. In Fig. 4, we can see how the databases we used to find these papers were distributed. The majority of the publications were found in the Springer, Google Scholar, and Scopus databases, as shown in Table 1. The four research issues have been addressed by extracting and synthesizing data from the chosen studies. The primary goals of the data retrieval process were to answer the research questions and verify that the studies fulfilled the exclusion criteria. It is in Appendix A that you will find the studies that were chosen once they met the criteria for exclusion. The data synthesis process included combining and synthesizing all of the gathered data in order to provide answers to the study questions.

Results

Table 2 displays the chosen publications. The year of publication, title, and algorithms utilized in these articles are shown in the table. The annual publishing rate over the last decade is shown in Figure 4. This data shows that there has been a surge in the number of publications

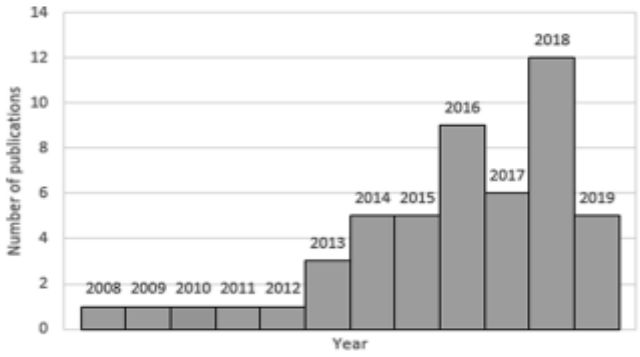

Fig. 4. Distribution of the selected publications per year.

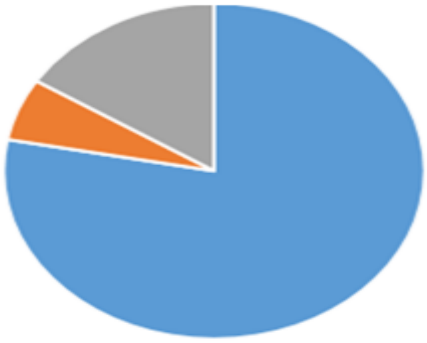

Fig. 5. Distribution of the type of 50 primary publications

Papers presented at conferences were also included since no specific kind of publishing was deemed inappropriate. Various kinds of publications are shown in the pie chart in Figure 5. As you can see from the image, the majority of the publications that were accessed were journal articles. Less than 25% of the total papers were conference papers or book chapters. Research question two (RQ2) was addressed by investigating and summarizing the characteristics employed in the machine learning techniques used in the publications. Table 3 displays all of the characteristics that could be extracted. The most popular characteristics are those that relate to weather, including temperature, rainfall, and soil type (Table 3). The crop yield is the one that is being studied. Grouping the characteristics allowed for a more comprehensive picture of the independent factors. Soil and crop data, humidity, nutritional levels, and field management are the categories into which the individual characteristics fall. In Table 4, you can see how often these categories are utilized. The most popular feature groups are those that pertain to soil, sun, and humidity data, as seen in the table. The variables that make up the feature group "soil information" include production area, pH value, soil type, soil maps, and cation exchange capacity. No two publications agree on whether or not soil maps were used, and if so, which ones. You may get broad details on the soil's nutrients, kind, and location in the soil maps.

Details on the crop itself, including its density, variety of plants, weight, and growth rate throughout the growing phase are known as crop information. The leaf area index and other growth indicators fall under this category as well. To represent the water on the field, we have humidity. Precipitation, predicted precipitation, humidity, and rainfall all belong to the humidity category. There are two types of nutrients: those that are already present in the soil and those that are added. These characteristics quantify the saturation level. The elements that have been measured include phosphorus, magnesium, potassium, sulfur, zinc, boron, calcium, and manganese. Decisions made by farmers to modify their fields are consolidated via field

management. Field management might also include nutrient management as these elements include irrigation and fertilization. Solar data includes characteristics related to heat and radiation. Shortwave radiation, degree-days, temperature, photoperiod, gamma radiometric, and solar radiation are among these. Features that do not fit into any of the other categories are housed in the "Other" feature group. According to Measuring Vegetation (NDVI & EVI), 2000, the majority of these features are either computed or utilized just once. characteristics like wind speed, pressure, and pictures fall under this category of characteristics that get little usage. Enhanced Vegetation Index (EDVI), Nor-Table 2 Malized Vegetation Index (NDVI), and MODIS Enhanced Vegetation Index (MODIS-EVI) are the computed characteristics.

Table 3
All features used.

Feature	# of times used
Temperature	24
Soil type	17
Rainfall	17
Crop information	13
Soil maps	12
Humidity	11
pH-value	11
Solar radiation	10
Precipitation	9
Images	8
Area of production	8
Fertilization	7
NDVI	6
Cation exchange capacity	6
Nitrogen	6
Irrigation	5
Potassium	5
Wind speed	5
Zinc	3
Magnesium	3
Shortwave radiation	2
Sulphur	2
Boron	2
Calcium	2
Organic carbon	2
EVI	2
Phosphorus	2
Gamma radiametrics	1
MODIS-EVI	1
Forecasted rainfall	1
Photoperiod	1
Climate	1
Degree-days	1
Time	1
Pressure	1
Leaf area index	1
Manganese	1

Group	# of times used
Soil information	54
Solar information	39
Humidity	38
Nutrients	28
Other	24
Crop information	14
Field management	12

Fig. 6 illustrates the important characteristics and sub-features shown in a feature map that we created to represent all the features collected in this SLR investigation. We looked into and summarized machine learning algorithms to

answer the first research question (RQ1). You can find the algorithms that were used several times in Table 5. The two most common algorithms, as shown in the table, are linear regression and neural networks (NN). Table 5 also shows the widespread usage of Random Forest (RF) and Support Vector Machines (SVM). Research question three (RQ3) was addressed by identifying assessment parameters. Table 6 displays all the assessment parameters and their frequency of application. The chart clearly illustrates that the most often utilized parameter in these investigations is the Root Mean Square Error (RMSE). Multiple validation approaches were also used in addition to the assessment parameters. The technique of cross-validation is often used. For this examination, 10-fold cross-validation was the technique of choice. Reading the articles to identify any issues or suggestions for future models helped answer research question four (RQ4). A lack of data or inadequate data was identified as an issue in many research. The tests demonstrated that their methods were effective with the data they had, but they recommended using more diverse data in future experiments. This includes data from a variety of climates and plant types, as well as data from lengthier time series of yields. The integration of other data sources is another area that may need enhancement. The article concluded by recommending more research into the use of machine learning to agricultural management software. It is necessary to develop software programs that enable the farmer to make choices using the models, provided that the models function as expected.

Conclusion

This study showed that the selected publications use a variety of features, depending on the scope of the research and the availability of data. Every paper investigates yield prediction with machine learning but differs from the features. The studies also differ in scale, geological position, and crop. The choice of features is dependent on the availability of the dataset and the aim of the research. Studies also stated that models with more features did not always provide the best performance for the yield prediction. To find the best performing model, models with more and fewer features should be tested. Many algorithms have been used in different studies. The results show that no specific conclusion can be drawn as to what the best model is, but they clearly show that some machine learning models are used more than the others. The most used models are the random forest, neural networks, linear regression, and gradient boosting tree. Most of the studies used a variety of machine learning models to test which model had the best prediction. Since Neural Networks is the most applied algorithm, we also aimed to investigate to what extent deep learning algorithms were used for crop yield prediction. After the identification of 30 papers that applied deep learning, we extracted and synthesized the applied algorithms. We observed that CNN, LSTM, and DNN algorithms are the most preferred deep learning algorithms. However, there are also other kinds of algorithms applied to this problem. We consider that this article will pave the way for further research on the development of crop yield prediction problem. In our future work, we aim to build on the outcomes of this study and focus on the development of a DL-based crop yield prediction model.

References

[1]. Ahamed, A.T.M.S., Mahmood, N.T., Hossain, N., Kabir, M.T., Das, K., Rahman, F., Rahman, R.M., 2015. Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh. In: 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2015 - Proceedings, https://doi.org/10.1109/SNPD.2015.7176185.

[2]. Ahmad, I., Saeed, U., Fahad, M., Ullah, A., Habib-ur-Rahman, M., Ahmad, A., Judge, J., 2018. Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan. J. Indian Soc. Remote Sens. 46 (10), 1701–1711. https://doi.org/10.1007/s12524-018-0825-8.

[3]. Ali, I., Cawkwell, F., Dwyer, E., Green, S., 2017. Modeling managed grassland biomass estimation by using multitemporal remote sensing data—a machine learning approach. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10 (7), 3254–3264. https:// doi.org/10.1109/JSTARS.2016.2561618.

[4]. Alpaydin, E., 2010. Introduction to Machine Learning, 2nd ed. Retrieved from https:// books.google.nl/books?hl=nl&lr=&id=TtrxCwAAQBAJ&oi=fnd&pg=PR7&dq= introduction+to+machine+learning&ots=T5ejQG_7pZ&sig=0xC_

H0agN7mPhYW7oQsWiMVwRnQ#v=onepage&q=introduction to machine learning&f=false. Ananthara,

[5]. M.G., Arunkumar, T., Hemavathy, R., 2013. CRY-An improved crop yield prediction model using bee hive clustering approach for agricultural data sets. In: Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, PRIME 2013, 473–478. https://doi.org/10.1109/ICPRIME.2013.6496717. Ayodele, T.O., 2010. Introduction to Machine Learning.

[6]. Baldi, P., 2012. Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of ICML workshop on unsupervised and transfer learning, pp. 37–49.

[7]. Baral, S., Kumar Tripathy, A., Bijayasingh, P., 2011. Yield Prediction Using Artificial Neural Networks, pp. 315–317. https://doi.org/10.1007/978-3-642-19542-6_57. Bargoti, S., Underwood, J.P., 2017. Image segmentation for fruit detection and yield estimation in apple orchards.

[8]. J. Field Rob. 34 (6), 1039–1060. https://doi.org/10.1002/rob.21699. Beulah, R., 2019. A survey on different data mining techniques for crop yield prediction. Int. J. Comput. Sci. Eng. 7 (1), 738–744. https://doi.org/10.26438/ijcse/v7i1.738744.

[9]. Bhojani, S.H., Bhatt, N., 2020. Wheat crop yield prediction using new activation functions in neural network. Neural Comput. Appl. 1–11. Bose, P., Kasabov, N., Bruzzone, L., n.d. Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series. Ieeexplore.Ieee.Org. Retrieved from https://ieeexplore.ieee.org/abstract/document/7524771/.

[10]. Brownlee, J., 2016. Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras. Machine Learning Mastery. Brownlee, J., 2017. Long Short-term Memory Networks with Python: Develop Sequence Prediction Models with Deep Learning. Machine Learning

[11]. Mastery. Brownlee, J., 2019. Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python. Machine Learning Mastery. Cakir, Y., Kirci, M., Gunes, E.O., 2014. Yield prediction of wheat in south-east region of Turkey by using artificial neural networks. In: 2014 The 3rd International Conference on Agro-Geoinformatics, Agro-Geoinformatics 2014. https://doi.org/10.1109/AgroGeoinformatics.2014.6910609.