KCG
COLLEGE OF TECHNOLOGY

KCG COLLEGE OF TECHNOLOGY

KARAPAKKAM

CHENNAI

LITERATURE SURVEY

TOPIC: ESTIMATE THE CROP YIELD USING DATA ANALYTICS

TEAM MEMBERS :

VIMAL KHANNA M

YOGESH J

MOHAMED FAYAZ S

VISHNUKANTH M S

Crop production in India is one of the most important sources of income and India is one of the top countries to produce crops. As per this project we will be analyzing some important visualization, creating a dashboard and by going through these we will get most of the insights of Crop production in India.

CROP YIELD PREDICTION USING MACHINE LEARNING: A SYSTEMATIC LITERATURE REVIEW

Publisher: computers-and-electronics-in-agriculture

Author: Thomasvan Klompenburg,[a]AyalewKassahun,[a]CagatayCatal[b]

Crop yield prediction is an essential task for the decision-makers at national and regional levels (e.g., the EU level) for rapid decision-making. 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.

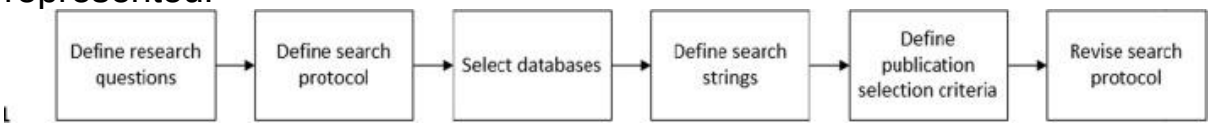
During our analysis of the retrieved publications, one of the exclusion criteria is that the publication is a survey or traditional review paper. Those excluded publications are, in fact, related work and are discussed in this section. Chlingaryan and Sukkarieh performed a review study on nitrogen status 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, 2016a, Gandhi and Armstrong, 2016b). 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.

3.1. Review protocol

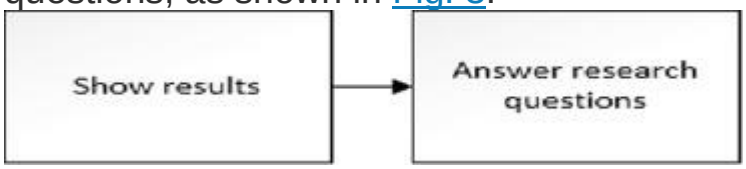
Before conducting the systematic review, a review protocol is defined. The review has been done using the well-known review guidelines provided by Kitchenham et al. (2007). Firstly, the research questions are defined. When research questions are ready, databases are used to select the relevant studies. 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.



3.2. Research questions

This SLR aims to get insight into what studies have been published in the domain of ML and crop yield prediction. To get insight, studies have been analyzed from several dimensions. For this SLR study, the following four research questions(RQs) have been defined.

•

RQ1- Which machine learning algorithms have been used in the literature for crop yield prediction?

- RQ2- Which features have been used in literature for crop yield prediction using machine learning?

- RQ3- Which evaluation parameters and evaluation approaches have been used in literature for crop yield prediction?

- RQ4- What are challenges in the field of crop yield prediction using machine learning?

3.3. Search strategy

The searching is done by narrowing down to the basic concepts that are relevant for the scope of this review. Machine learning has many application fields, which means that there are a lot of published studies that are probably not in the scope of this review article. The basic searching is done by an automated search. The starting input for the search was “machine learning” AND “yield prediction”. Articles were retrieved, and abstracts were read to find the synonyms of the keywords. The search was performed in six databases. The search input “machine learning” AND “yield prediction” was used to get a broad view of the studies. After the exclusion criteria were applied, and all the results were processed, and a more complex search string was built in order to avoid missing relevant studies. This final search string is as follows: ((“machine learning” OR “artificial intelligence”) AND “data mining” AND (“yield prediction” OR “yield forecasting” OR “yield estimation”)). After executing this search string, 567 studies were retrieved.

A specific description of the search strings per database are provided as follows:

Science direct: The search string is [“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, keywords).
Scopus: The search string is [“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, keywords).

Web of Science: The search string is [“machine learning” AND “yield prediction”] (title, abstract, author keywords, and Keywords Plus).
Springer Link: The search string is [“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: The search string is [“machine learning” AND “yield prediction”] (anywhere).
Google Scholar: The search string is [“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).
For Web of Science and Wiley, the search string [((“machine learning” OR “artificial intelligence”) AND “data mining” AND (“yield prediction” OR “yield forecasting” OR “yield estimation”))] did not result in any publications.

3.4. Exclusion criteria

To exclude irrelevant studies, the studies were analyzed and graded based on exclusion criteria to set the boundaries for the systematic review. The exclusion criteria (EC) are shown as follows:

Exclusion criteria 1 - Publication is not related to the agricultural sector and yield prediction combined with machine learning
Exclusion criteria 2 – Publication is not written in English
Exclusion criteria 3 – Publication that is a duplicate or already retrieved from another database
Exclusion criteria 4 – Full text of the publication is not available
Exclusion criteria 5 – Publication is a review/survey paper
Exclusion criteria 6 – Publication has been published before 2008

After the first three exclusion criteria were applied, only 77 studies remained for further analysis. After applying all the six exclusion criteria, 50 studies were selected for further analysis. In Table 1, we show the number of initially retrieved papers and the number of papers after selection criteria were applied. Fig. 4 shows the distribution of selected publications based on the databases we searched. As shown in Table 1, most of the papers were retrieved from Google Scholar, Scopus, and Springer 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