Machine learning technology for early prediction of grain yield at the field scale: a systematic review

Joerg Leukel[a], Tobias Zimpel[a] and Christoph Stumpe[b]

[a]*Department of Information Systems 2, University of Hohenheim*

[b]*Department of Fundamentals of Agricultural Engineering, University of Hohenheim*

{joerg.leukel | tobias.zimpel | christoph.zumpe}@uni-hohenheim.de

Abstract. Machine learning (ML) has become an important technology for the development of prediction models for crop yield. Predictive modeling using ML is rapidly growing as research addresses early predictions versus predictions shortly before harvest, predictions at the scale of field or region, and predictions for different types of crops. This great diversity of prediction tasks requires a proper choice of specific ML techniques to attain high levels of performance. Therefore, this review focuses on a distinct prediction task and aims to provide task-specific insights into the adoption of ML. The objective of our research is to investigate ML approaches for the early prediction of grain yield at the field scale. We identified studies published between 2014 and 2021 through a systematic search in Scopus and Web of Science for journal articles and a retrieval of analogous articles from three previous reviews. The study selection process included screening, full-text assessment, and data extraction by two independent coders. Of 924 unique records identified in the search and retrieval, 157 full texts were assessed for eligibility, and 46 studies met all inclusion criteria. The results paint a comprehensive picture of the ML techniques used, revealing the richness of data collection, preprocessing, model training, and model evaluation. Specifically, the results highlight (1) a wide range of prediction horizons from a few weeks up to more than eight months before harvest; (2) a large set of input data representing weather, crop management, site characteristics, and vegetation properties; (3) a low level of adoption of feature selection methods to enhance performance; (4) some lack of information on the size of the training and test sets required to assess their suitability; and (5) heterogeneity in the reporting of performance metrics that hinders the comparison and integration of evidence from individual studies. To overcome barriers to the accumulation of evidence, we suggest recommendations for enhanced reporting and building greater consensus regarding the most appropriate performance metrics.

Keywords: Crop management; yield prediction; machine learning; precision agriculture; remote sensing.

This is the author accepted manuscript (AAM) of the following article:

Leukel, J., Zimpel, T., & Stumpe, C. (2023). Machine learning technology for early prediction of grain yield at the field scale: A systematic review. *Computers and Electronics in Agriculture*, 207, 107721. https://doi.org/10.1016/j.compag.2023.107721

1. Introduction

Yield prediction has an important role in crop farming aimed at efficient and sustainable production. Accurate and timely predictions are important for farmers' decision making regarding planting, irrigation, fertilization, harvesting, and trading. For the development of prediction models, machine learning (ML) has become a key technology. The principal idea of ML is to learn a prediction model from past data, evaluate the model based on new observations, and ultimately deploy the model in a productive environment (Jordan and Mitchell, 2015). The number of applications of ML technology for crop yield prediction have increased rapidly in the past few years. This growth has been amplified by freely available ML algorithms, improved remote sensing techniques, and the enhanced provision of smart farming data that represents genotypes, soil, weather, crop management, and other environmental parameters that affect crop growth (Wolfert et al., 2017).

The field of ML-based yield prediction has made great strides in enhancing the accuracy and robustness of prediction models. Evidence for the effectiveness of ML-based yield prediction can be categorized along at least three dimensions: (1) prediction horizon, ranging from a few hours to many months before harvest; (2) scale, such as field and region; and (3) type of crops, including grains and fruits in orchards. These dimensions span a large array of prediction tasks. At the same time, a great variety of ML techniques are available to researchers and practitioners who want to develop a prediction model. Developers must consider alternative techniques in each phase of development, such as the preprocessing of raw data into features, training from past time-series data, and the evaluation of performance on new data. The choices of these techniques have pivotal impacts on the usefulness of the prediction model. Therefore, gaining insights into ML techniques is essential for the understanding of the best practices through which high levels of performance can be attained.

With respect to the prediction horizon, research has demonstrated that predictions are feasible at all vegetation stages. For instance, some studies compared models for pre-sowing, mid-season and late-season predictions of grain yield and found better performance during the later stages (Fieuzal et al., 2020; Filippi et al., 2019). Similarly, a literature review by Muruganantham et al. (2022) noted that the relationship between vegetation indices obtained from images and crop yield is not static but varies by vegetation stage. Consistent with this conclusion, a review of 69 studies by Benos et al. (2021) highlighted a handful of articles that examined predictions at a specific vegetation stage or time before harvest. Notwithstanding these indications for the importance of the prediction horizon, many studies focused on predictions right before harvest. However, the results of studies on early prediction cannot necessarily be compared with the results of studies on predictions right before harvest when much more within-season data is available. The conceptual and empirical differences of early prediction warrant the further assessment of the evidence from previous studies.

Scale is an important dimension of prediction models because models at each scale serve a different purpose. While plant-scale models are aimed to better understand the factors affecting crop growth (Shekoofa et al., 2014), field-scale models can directly assist in crop management

(Basso and Liu, 2019) and models at larger scales primarily inform policy making in agriculture (López-Lozano et al., 2015). Depending on the scale, models often largely differ in the scope, amount and granularity of the agricultural input data used, which then impacts the adoption of ML techniques. A variety of scales has been identified in previous systematic reviews (van Klompenburg et al., 2020). Specifically, a recent review of 44 studies found that most models predicted crop yield at the regional scale rather than the field scale (Oikonomidis et al., 2022). The review also accentuated the higher effort required for collecting field-level yield data (e.g., in-field sampling) compared with accessing readily available yield data from governmental bodies and regional associations. Although prediction models at the regional scale can exhibit good accuracy, their usefulness to inform the decision making of individual farmers might be severely limited. Taken together, these results suggest that increased attention should be given to different scales when assessing the role of specific ML techniques for achieving high levels of prediction performance.

A third concern in the literature is the development of prediction models that are specific to one crop or type of crop. Although the digital technologies available for remote sensing, image processing, model training and evaluation are independent from the surveyed crop, the vast majority of studies have only focused on one crop. The rationale is that differences in crop phenology and cultivation patterns affect the spatial and temporal variability of input data for training a prediction model. These differences concern major types of crops, such as grains, fruit crops, and root vegetables. For this reason, crop-specific reviews have been undertaken for fruits in orchards (He et al., 2022), vineyards (Barriguinha et al., 2021), palm oil plantations (Rashid et al., 2021), and pasture crops (Morais et al., 2021). Further reviews considered a wider range of crops but identified the specific studied crops (Benos et al., 2021; Oikonomidis et al., 2022; van Klompenburg et al., 2020).

Previous reviews of the literature have made important contributions to understanding the adoption of agricultural input data, general ML algorithms, and performance metrics for a large array of prediction tasks (Bali and Singla, 2022; Benos et al., 2021; van Klompenburg et al., 2020). In light of the three dimensions of yield prediction discussed above, providing further insights is possible by focusing on a specific prediction task. Against this backdrop, our research investigates ML technology for the early prediction of grain yield at the field scale. The focus on the early prediction is guided by the hypothesis that prediction performance is negatively correlated with prediction horizon; hence, we intend to validate this hypothesis through our review. By focusing on the field scale, we expect to gain task-specific insights into the adoption of ML techniques. We chose grain as the type of crop under study to further reduce variability in the prediction models. This approach facilitates the collation and synthesis of quantitative results from otherwise heterogeneous studies. Specifically, the objective of our research is to assess the adoption of ML technology for the early prediction of grain yield at the field scale by conducting a systematic review of published studies.

2. Method

This systematic review was conducted in accordance with guidelines defined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement where applicable (Moher et al., 2009). In this section, we discuss the criteria for the inclusion of studies, the information sources that were searched, and the processes for study selection, data extraction, and quantitative data synthesis. To minimize error and bias, two authors coded the data and worked independently in all stages of the review (from study identification to data extraction).

2.1. Eligibility criteria

Studies were eligible for inclusion if they applied machine learning technology to predict grain yield at the field scale using real-world data such that there was a meaningful offset between the last recording of input data and harvest (so-called early prediction). We excluded studies for the following reasons: (1) yield prediction not related to grain but a different type of crop (e.g., fruits) or object (e.g., photovoltaic plant), (2) prediction of a non-continuous variable (e.g., growth stage), (3) prediction at a different scale (e.g., region), (4) prediction right before harvest (no offset), (5) no test of a prediction model but conceptual research or literature review, (6) no application of ML technology, and (7) and no real data set but synthetic data. With respect to bibliographic characteristics, we defined the following eligibility criteria: article published in a journal, written in English, and original contribution.

2.2. Information sources and search

The identification of articles relied upon a systematic search of journal articles published between 2014 and 2021. We carried out the search using the electronic databases Scopus (January 2022) and Web of Science (November 2022), which have comprehensive coverage of peer-reviewed articles. The automated bibliographic search was complemented with analogous articles included in three previous systematic reviews of ML-based yield prediction (Bali and Singla, 2022; Benos et al., 2021; van Klompenburg et al., 2020).

The bibliographic search was performed on each article's title, abstract, and keywords by combining search terms for crop yield prediction and ML. The prediction task was represented as ("crop predict*" OR "crop forecast*" OR "crop estimat*" OR "yield predict*" OR "yield estimat*" OR "yield forecast*") to account for different terminology in the literature. The coding of ML both covered abstract terms and concrete ML algorithms by representing ML as ("machine learning" OR "deep learning" OR "artificial intelligence" OR "support vector" OR "random forest*" OR "neural network*" OR ANN OR SVM). To prevent the oversight of articles that provide no specification of the grain and scale of prediction in their title, abstract, or keywords, we included neither grain nor field scale in the search query.

2.3. Study selection

After the removal of duplicate records, two authors independently carried out the screeningwith a codebook describing the eligibility and exclusion criteria. All codes were compared, and disagreements were resolved by discussion. For the articles that passed the screening phase, one author downloaded the full texts from the publishers. The same coders independently assessed the full texts using the same codebook as in the screening phase. Finally, the codes were compared, and any inconsistencies were resolved by discussion between the coders. The initial agreement between coders in the screening phase (89.4%) and eligibility assessment (82.2%) was high.

2.4. Data collection process

For the articles that went through the full-text assessment, the two coders independently extracted data using a codebook for the data items defined in Section 2.5 (data points were recorded in a standardized spreadsheet form). Afterwards, the coders discussed all individual codes to agree upon the final data points. The agreement between coders was high (93.8%).

2.5. Data items

The conceptual model of the review is shown in Fig. 1. This model structures the process for ML-based grain yield prediction and indicates the major data items.

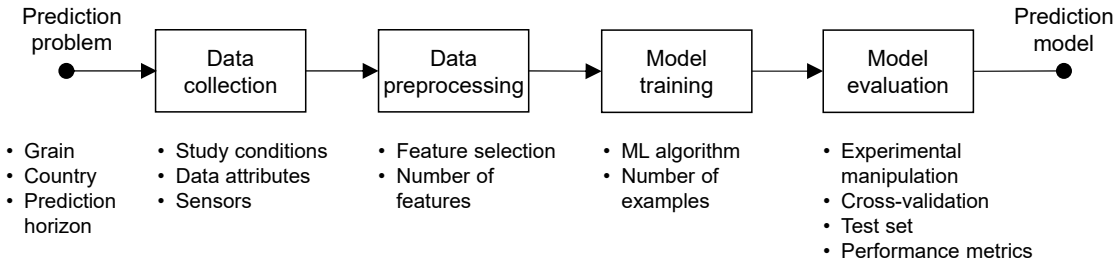

Fig. 1. Machine learning process and data items for grain yield prediction.

Prediction is used to forecast the yield of different grain crops in the future. Because the yield will be known at harvest, the predicted yield can eventually be compared to the observed yield. *Grain* is the crop that is cultivated and for which the prediction is made. *Country* describes the area where the study is conducted. *Prediction horizon* represents how far ahead the predictions are made, which can range from a handful of days up to many months before harvest. As shown in Fig. 1, the process for ML-based yield prediction is defined by sequential phases of data collection, data preprocessing, model training, and model evaluation, which we discuss in the following paragraphs.

Data collection includes the acquisition of prediction-relevant data and the creation of a data set. *Study conditions* define the fields, seasons, and genotypes for which data are collected during crop growth. Many studies organize the field into a larger number of plots; data are collected for each plot and the plot-scale yield predictions are converted into field-scale predictions. *Data attributes* are the different input data that are collected for the field (or plot) under study. Attributes

can be categorized as follows: weather (e.g., precipitation and temperature), crop management (e.g., fertilization and irrigation), site (e.g., soil properties); vegetation indices obtained from spectroradiometers (Xue and Su, 2017), and canopy (e.g., height above the ground). Attributes of the two latter categories are measured using *sensors* that can be classified by the distance from the ground as follows (Mouazen et al., 2020). First, satellites observe a crop from orbit through optical sensors, which allow for the interpretation of reflectance with respect to vegetation stages (Zeng et al., 2020). Second, unmanned aerial vehicles (UAVs) carry devices, such as hyperspectral and multispectral cameras, and survey a field from a few meters above to provide high-resolution imagery (Olson and Anderson, 2021). Third, in-field measurement spans from the use of handheld sensors and stand assessment (visual inspection) to the destructive sampling of plants for subsequent laboratory analysis.

Data preprocessing is the phase of transforming the data set into a representation from which a prediction model can be trained. Although preprocessing often includes many laborious subtasks, such as the integration of different raw formats and the handling of incomplete data, these subtasks can be rather easily solved (Raschka, 2015). Therefore, we focus on *feature selection* for defining the best subset of features from attributes. The specific feature selection methods can be categorized as follows (Chandrashekar and Sahin, 2014). Filter-based methods calculate a metric and select features based on that metric (e.g., correlation coefficient). Wrapper-based methods remove different features from a subset, evaluate the goodness of the subset, and eventually choose the subset with the best evaluation; example methods are backward elimination and random choice. Embedded methods are built-in specific ML algorithms, such as Random Forests feature selection. The *number of features* denotes the total number of features that are forwarded to the training phase.

Model training is concerned with learning a function that best maps an input onto an output based on example input-output pairs. In crop yield prediction, the input includes all the selected features at the time of prediction, and the output is the observed yield at harvest. *ML algorithm* refers to the supervised learning algorithm used to estimate the mapping function, such as Artificial Neural Networks (ANN) (Bishop, 2006) and Random Forests (RF) (Ho, 1995). Regardless of the specific algorithm used, a sufficiently large *number of examples*, or pairs of inputs and observed yields, are required.

Model evaluation is used to assess the performance of a trained prediction model. Evaluation usually relies upon the *experimental manipulation* of one or more factors. Manipulation allows for the determination of effects on performance and the discovery of experimental conditions that enhance performance. In evaluation, the trained model must be tested on new observations, i.e., data that were not used for training the model (also called unknown data). One technique is *cross-validation*, which partitions an entire data set into complementary subsets (denoted as *k*-folds), conducts training on *k*-1 subsets, and eventually tests the prediction model on the remaining subset. Performance is then calculated as the mean of the results for all of the *k*-folds. Another technique tests a prediction model on a separate data set of new observations (so-called *test set*). For either technique, new observations for testing the model can be defined temporally (i.e., observations in a different season) or cross-sectionally (i.e., observations in the same season). The

prediction model is evaluated using *performance metrics*, which are available for comparing a set of predicted yields with a corresponding set of observed yields. Example metrics are the coefficient of determination (denoted as R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

2.6. Quantitative data synthesis

We analyzed the relationship between prediction horizon and performance and focused on the R^2 metric because it is unitless and thus enabled collation of performance across studies. For each study that reported R^2, we first checked if the prediction horizon in months was available. If this information was not available from the article, we contacted the corresponding author. We excluded a study if it used no specific prediction horizon but trained the model by pooling data with different timespans between the last input and harvest. If a study reported separate results for different horizons, we selected the models with the smallest and largest horizons.

In an additional analysis, we took differences in growth periods into account. For instance, if the growth period lasts only three months, a prediction horizon of two months might be regarded as very early. However, this interpretation depends on the length of the growing period, and thus it could be different for other periods. For each prediction model, we calculated the prediction time relative to the growth period (defined as the time from sowing to prediction divided by the length of the growth period). This measure indicates how far along the growth period is. It ranges from 0 for prediction at sowing to 1 for prediction at harvest. Complete information on the extraction of prediction horizons, R^2, and growth periods from the studies is given in the Supplement.

3. Results

3.1. Study selection

Fig. 2 presents the selection process in a PRISMA flow diagram and indicates the reasons for the exclusion of records and articles. We retrieved 736 records from Scopus, 751 records from Web of Science, and 133 records from three previous reviews. After the removal of duplicate records, a total of 924 records were screened based on the title, abstract, and list of keywords. Of these, 157 articles were selected for the full-text assessment, and 46 studies met the criteria for inclusion.

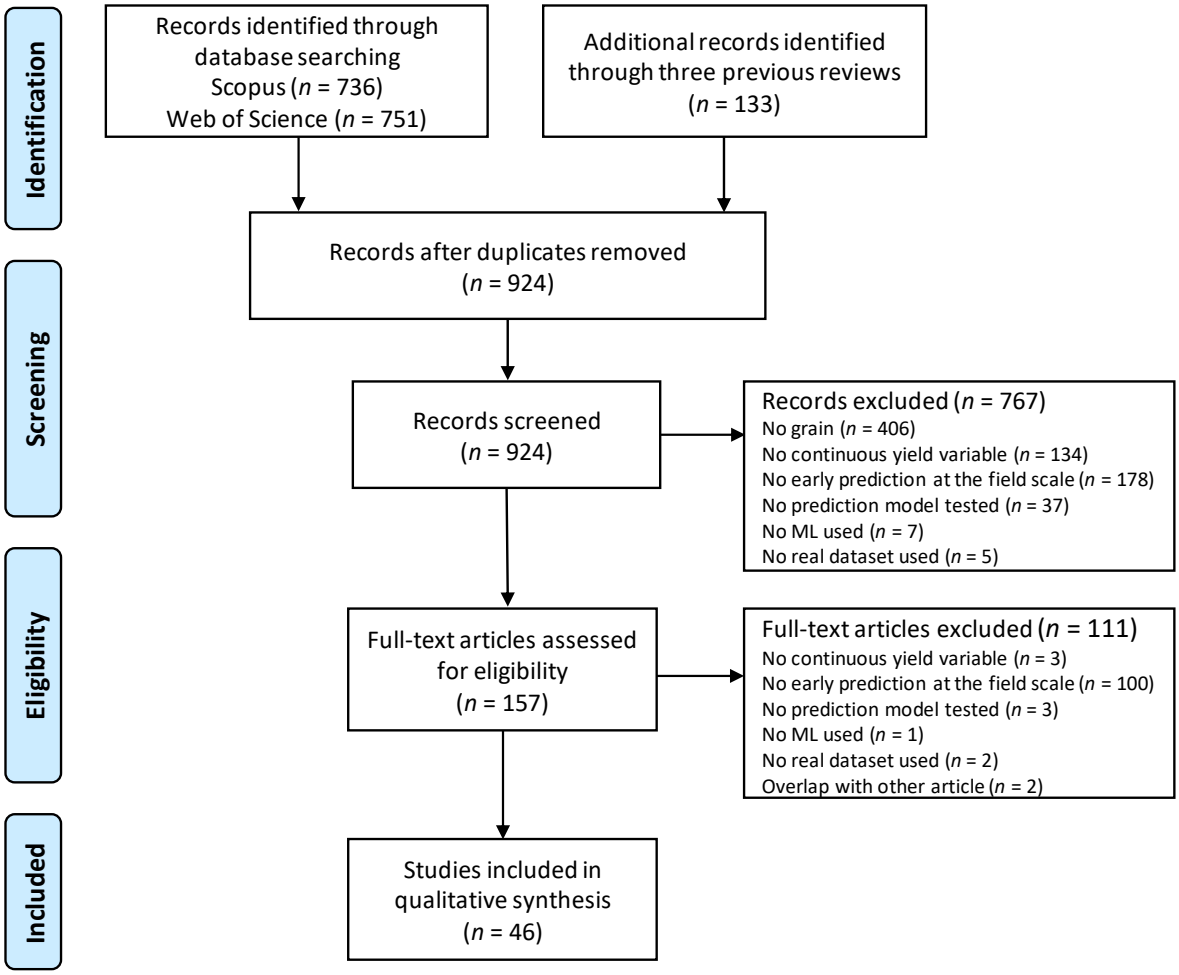

Fig. 2. Flow diagram of study selection.

Table 1 provides an overview of the studies, the majority of which were published in 2021 (23) and 2020 (12). The most frequently studied grains were wheat (22), maize (13), soybean (5), barley (4), rapeseed (4), and rice (4). Eleven studies were conducted in China, and nine studies were conducted in the United States of America. For a total of 38 studies, we were able to identify the time period between prediction and harvest measured in months (depending on the unit and details reported). Overall, the time periods ranged from more or less one month up to more than eight months. For five of the remaining studies, the reporting lacked sufficient details to identify the prediction

horizon and was limited to developmental stages of the crop (e.g., tillering, stem extension, and heading). No such information was available for three studies.

Table 1. List of studies included in qualitative synthesis ($N = 46$).

Study	Grain	Country	Prediction horizon
Adak et al. (2021)	Maize	USA	0.9 m
Alebele et al. (2021)	Rice	China	1 to 2.5 m (*)
Barbosa et al. (2020)	Maize	USA	NR
Barzin et al. (2020)	Maize	USA	3 to 5 m (*)
Basir et al. (2021)	Rice	Bangladesh	4 m
Castaldi et al. (2015)	Wheat	Italy	0 to 5 m (*)
Chen and Jing (2017)	Wheat	China	1.5 m
Choudhury et al. (2021)	Wheat	Australia	1.4 m
Costa et al. (2022)	Wheat	USA	0.5, 0.7 m
Danilevicz et al. (2021)	Maize	USA	3.3 m
da Silva et al. (2020)	Soybean	Brazil	1 growth stage
Eugenio et al. (2020)	Soybean	Brazil	8 m (*)
Fajardo and Whelan (2021)	Wheat	Australia	3 m
Fan et al. (2021)	Rapeseed	China	1.2 m (*)
Fei et al. (2021a)	Wheat	China	1 growth stage
Fei et al. (2021b)	Wheat	China	6 growth stages
Feng et al. (2020)	Wheat	Australia	6 growth stages
Fieuzal et al. (2017)	Maize	France	0.1 to 4.6 m (*)
Fieuzal et al. (2020)	Wheat	France	0 to 8 m
Filippi et al. (2019)	Wheat; barley; rapeseed	Australia	1.5 to 7 m (*)
Florence et al. (2021)	Wheat	UK (Scotland)	1.7 to 4.7 m
García-Martínez et al. (2020)	Maize	Mexico	4, 5 m
Habyarimana and Baloch (2021)	Sorghum	Italy	0 to 7 m (*)
Hassanzadeh et al. (2021)	Snap bean	USA	0 to 0.8 m
Hunt et al. (2019)	Wheat	UK (England)	1.2 to 8.3 m (*)
Kross et al. (2020)	Maize; soybean	Canada	1 to 2 m (*)
Li et al. (2021)	Wheat; maize; rice	China	6 growth stages
Meng et al. (2021)	Maize	USA	1 to 2 m (*)
Nevavuori et al. (2019)	Wheat; barley	Finland	1, 2, 3 m (*)
Nevavuori et al. (2020)	Wheat; barley; oat	Finland	1.8 to 3.1 m (*)
Ngie and Ahmed (2018)	Maize	South Africa	1, 3 m (*)
Niedbała et al. (2019a)	Wheat	Poland	1.4 to 3.7 m (*)
Niedbała et al. (2019b)	Rapeseed	Poland	0.5 to 2.8 m (*)
Ozcan et al. (2021)	Wheat	Turkey	1 m
Ramos et al. (2020)	Maize	Brazil	NR
Sagan et al. (2021)	Maize; soybean	USA	0.8 to 1.2 m (*)
Šestak et al. (2018)	Wheat	Croatia	2.2 m
Shafiee et al. (2021)	Wheat	Norway	0.6 to 1.6 m (*)
Sharifi (2021)	Barley	Iran	1 m
Tian et al. (2021)	Wheat	China	0.7 m
Wan et al. (2020)	Rice	China	0 to 3 m
Wen et al. (2021)	Rapeseed	Canada	2 m (*)
Zhang et al. (2020)	Wheat	China	NR
Zhang et al. (2021)	Maize	China	1 to 6 m (*)
Zhou et al. (2021a)	Soybean	USA	1.8 m (*)
Zhou et al. (2021b)	Wheat	Japan	1.0, 1.1 m

Note. m = months. (*) = no exact dates and timespans available.

3.2. Data collection

Table 2 shows the number of seasons and lists the data attributes per study. In most cases, data were collected in one (21) or two (9) seasons. Thirty-seven articles clearly specified whether the data had been collected at a research site (20) or a field operated by a farmer (17) (not tabulated). The results for the different categories of data attributes highlight the importance of vegetation indices (VIs) and management data, which were present in 38 and 25 of the data sets, respectively. Many studies assessed alternative or complementary VIs, with every fourth study testing at least eight different indices. Management attributes often included fertilizer input (14) and planting date (11), and the most incident weather attributes were precipitation (14) and temperature (13). Every fourth study collected site data, such as electroconductivity, elevation, soil moisture, and surface roughness. Every seventh study objectively measured canopy height in the field.

Table 3 provides information on the fields and types of sensors used. Seventeen studies collected the data of a single field, whereas five studies had access to data of hundreds of fields. The total field size ranged from less than one hectar for some experiments at research facilities to 3300 hectares, although most studies did not report that information. More than half of the studies investigated a single genotype, but a handful of studies considered hundreds of varieties.

Satellites and UAVs were the most frequently used sensors (21 studies each). Eleven studies analyzed data collected via two sensor types. For instance, one each study combined vegetation indices derived from satellite images with UAV-based laser scans (Kross et al., 2020) and RGB images (Sagan et al., 2021), respectively. Frequent sources of satellite images were Sentinel-2 (6) and Landsat (5), and UAVs mostly carried multispectral (16) and RGB (7) cameras. As the spatial resolution of images depends on the specific camera and satellite used, the resolution spanned from 1 to 10 centimeters for UAVs and 10 to 1000 meters for satellites, respectively. A broad range of in-field sensors was found by including the following: handheld sensors for canopy reflectance and vegetation indices; stand assessment by counting the number of plants (da Silva et al., 2020) and rating the canopy wilting (Zhou et al., 2021a); and destructive sampling for determining leaf area index (Fan et al., 2021) and soil roughness (Fieuzal et al., 2017). Three studies did not collect data via specific sensors but exclusively retrieved readily available data about weather, soil, or crop management.

Table 2. Data collection in studies ($N = 46$).

Study	Seasons	Data attributes				
		Weather	Management	Site	VI	Canopy
Adak et al. (2021)	1	–	PD	–	12	Height
Alebele et al. (2021)	2	–	–	–	6	–
Barbosa et al. (2020)	1	–	FI, SR	Yes	–	–
Barzin et al. (2020)	3	PR	FI, PD	–	26	–
Basir et al. (2021)	1	–	SR	Yes	–	–
Castaldi et al. (2015)	1	–	–	–	2	–
Chen and Jing (2017)	1	–	–	–	*	–
Choudhury et al. (2021)	1	–	–	–	3	Height
Costa et al. (2022)	1	–	PD	–	–	–
Danilevicz et al. (2021)	3	–	FI	–	8	–
da Silva et al. (2020)	2	–	SR	–	12	–
Eugenio et al. (2020)	1	–	IR	–	10	–
Fajardo and Whelan (2021)	3	–	–	Yes	–	–
Fan et al. (2021)	2	–	FI	–	4	–
Fei et al. (2021a)	1	–	–	–	20	–
Fei et al. (2021b)	1	TE	IR	–	22	–
Feng et al. (2020)	10	PR, RA, TE, O	FI, PD, IR, O	Yes	1	–
Fieuzal et al. (2017)	1	PR	IR, PD, O	Yes	1	–
Fieuzal et al. (2020)	4	–	–	–	1	–
Filippi et al. (2019)	3	PR	PD	Yes	1	–
Florence et al. (2021)	2	–	FI	–	2	Height
García-Martínez et al. (2020)	1	–	FI	–	6	–
Habyarimana and Baloch (2021)	2	–	–	–	3	–
Hassanzadeh et al. (2021)	2	–	–	–	2	–
Hunt et al. (2019)	1	PR, TE	PD	Yes	5	–
Kross et al. (2020)	3	–	–	Yes	3	–
Li et al. (2021)	7	EV, PR, TE, O	PD	Yes	2	–
Meng et al. (2021)	14	PR, TE	FI	Yes	3	–
Nevavuori et al. (2019)	1	–	–	–	1	–
Nevavuori et al. (2020)	1	TE	–	–	*	–
Ngie and Ahmed (2018)	1	–	–	–	11	–
Niedbała et al. (2019a)	8	PR, TE	FI, PD, O	–	–	–
Niedbała et al. (2019b)	8	PR, TE	FI, PD, O	–	–	–
Ozcan et al. (2021)	1	PR, RA, TE, O	PD	–	4	–
Ramos et al. (2020)	2	–	–	–	33	–
Sagan et al. (2021)	1	–	FI, IR	–	18	–
Šestak et al. (2018)	1	–	FI	–	2	–
Shafiee et al. (2021)	1	–	–	–	3	–
Sharifi (2021)	5	PR, TE, O	–	–	2	–
Tian et al. (2021)	10	PR, TE	–	–	3	–
Wan et al. (2020)	2	–	FI	–	13	Height
Wen et al. (2021)	4	PR, TE	FI	Yes	2	Height
Zhang et al. (2020)	3	–	–	–	9	–
Zhang et al. (2021)	3	PR, TE, O	–	Yes	6	–
Zhou et al. (2021a)	1	–	–	–	3	Height
Zhou et al. (2021b)	2	–	–	–	7	Height
Count	n/a	16	25	12	40	7

Note. EV = evaporation. FI = fertilizer input. IR = irrigation. O = other. PD = planting date. PR = precipitation. RA = radiation. SR = seed rate or density. TE = temperature. VI = vegetation indices.
* = reflection bands but no specific VI used.

Table 3. Fields and sensors used in studies ($N = 46$).

Study	Fields			Type of sensors			Spatial resolution [m]
	No.	Size [ha]	Genotypes	Satellite	UAV	In-field	
Adak et al. (2021)	2	NR	100	–	Yes	–	NR
Alebele et al. (2021)	60	NR	1	Yes	–	–	10
Barbosa et al. (2020)	9	360	1	Yes	–	–	–
Barzin et al. (2020)	1	0.8	1	–	Yes	–	NR
Basir et al. (2021)	1	0.02	1	–	–	–	–
Castaldi et al. (2015)	7	219	4	Yes	–	–	10
Chen and Jing (2017)	36	NR	1	Yes	–	–	NR
Choudhury et al. (2021)	1	0.14	18	–	Yes	Yes	0.01 (UAV)
Costa et al. (2022)	1	0.08	40	–	Yes	–	NR
Danilevicz et al. (2021)	1	NR	1113	–	Yes	–	NR
da Silva et al. (2020)	3	NR	1	–	Yes	Yes	0.10
Eugenio et al. (2020)	1	NR	1	–	Yes	–	0.07
Fajardo and Whelan (2021)	11	3300	1	Yes	–	Yes	10
Fan et al. (2021)	3	NR	1	–	Yes	Yes	NR
Fei et al. (2021a)	1	NR	211	–	Yes	–	NR
Fei et al. (2021b)	1	NR	30	–	Yes	–	NR
Feng et al. (2020)	29	NR	>1	Yes	–	–	NR
Fieuzal et al. (2017)	30	0.5 to 39.3	1	Yes	–	Yes	8 to 20
Fieuzal et al. (2020)	12	3.2 to 28.6	1	Yes	–	–	10
Filippi et al. (2019)	NR	NR	>1	Yes	–	Yes	250
Florence et al. (2021)	1	0.10	2	–	–	Yes	–
García-Martínez et al. (2020)	1	0.24	1	–	Yes	–	0.02
Habyarimana and Baloch (2021)	23	174	>1	Yes	–	Yes	10
Hassanzadeh et al. (2021)	1	NR	6	–	Yes	–	0.03
Hunt et al. (2019)	39	662	1	Yes	–	–	10
Kross et al. (2020)	22	NR	2	Yes	Yes	–	1 (UAV)
Li et al. (2021)	220*	NR	6	Yes	–	–	NR
Meng et al. (2021)	1	28.8	1	Yes	–	–	NR
Nevavuori et al. (2019)	9	89.3	1	–	Yes	–	NR
Nevavuori et al. (2020)	9	85.1	1	–	Yes	–	NR
Ngie and Ahmed (2018)	2	208	NR	Yes	–	–	10
Niedbała et al. (2019a)	301	NR	1	–	–	–	–
Niedbała et al. (2019b)	328	NR	1	–	–	–	–
Ozcan et al. (2021)	142	NR	1	Yes	–	–	NR
Ramos et al. (2020)	1	NR	11	–	Yes	–	NR
Sagan et al. (2021)	3	1.34	99	Yes	Yes	–	0.01 (UAV)
Šestak et al. (2018)	1	3.9	1	–	–	Yes	–
Shafiee et al. (2021)	1	NR	394	–	Yes	–	NR
Sharifi (2021)	NR	NR	1	Yes	–	–	10
Tian et al. (2021)	10	NR	1	Yes	–	–	500
Wan et al. (2020)	1	NR	1	–	Yes	Yes	NR
Wen et al. (2021)	5	NR	3	–	–	Yes	–
Zhang et al. (2020)	NR	NR	>2	Yes	–	–	30
Zhang et al. (2021)	7531	NR	1	Yes	–	–	1000
Zhou et al. (2021a)	1	NR	116	–	Yes	Yes	0.01 (UAV)
Zhou et al. (2021b)	2	NR	1	–	Yes	–	0.06
Count	n/a	n/a	n/a	21	21	12	n/a

Note. NR = not reported. UAV = unmanned aerial vehicle. * = different per type of grain.

3.3. Data preprocessing

Table 4 indicates that 21 studies reported the adoption of a feature selection method. Filter-based methods included the use of correlation analysis to remove strongly correlated features (six studies) and the use of principal component analysis to transform strongly correlated features into a smaller number of principal components (three studies). The most frequent embedded method was Random Forests feature selection (Barzin et al., 2020; Ozcan et al., 2021; Ramos et al., 2020; Sagan et al., 2021). All other methods were adopted in one study each, including wrapper-based methods, such as particle swarm optimization (Hassanzadeh et al., 2021) and sequential forward selection (Shafiee et al., 2021). Regarding the number of selected features, information was only available in 20 articles (not tabulated). Studies at the lower end used three (Barzin et al., 2020; Ngie and Ahmed, 2018) and four features (Florence et al., 2021; Wen et al., 2021; Zhang et al., 2020). Eight studies considered at least 20 features, with a maximum of 35 features (Fajardo and Whelan, 2021).

Table 4. Types of feature selection methods in studies ($N = 46$).

Type	No. of studies	Studies
Filter-based feature selection	10	Chen and Jing (2017); Choudhury et al. (2021); Da Silva et al. (2020); Eugenio et al. (2020); Fei et al. (2021a; 2021b); García-Martínez et al. (2020); Ramos et al. (2020); Šestak et al. (2018); Zhang et al. (2020)
Embedded feature selection	8	Alebele et al. (2021); Barzin et al. (2020); Habyarimana and Baloch (2021); Ngie and Ahmed (2018); Ozcan et al. (2021); Ramos et al. (2020); Sagan et al. (2021); Zhang et al. (2021)
Wrapper-based feature selection	4	Feng et al. (2020); Hassanzadeh et al. (2021); Kross et al. (2020); Shafiee et al. (2021)

3.4. Model training

Table 5 shows that 18 different algorithms were applied in the selected studies. The most frequent algorithms were Artificial Neural Networks (20), Linear Regression (19), and Random Forests (18). Five other algorithms were only considered in one study each. The largest number of algorithms used in a single study was seven. Eleven studies tested two algorithms and twenty studies focused on one algorithm each.

Twenty-five articles report the size of the training set, by stating either the number of examples or a percentage value from which the number could be calculated (not shown in Table 5). The number varied between 16 in the study by Habyarimana and Baloch (2021) and more than 4000 in three studies (Danilevicz et al., 2021; Fajardo and Whelan, 2021; Hunt et al., 2019).

Table 5. Machine learning algorithms in studies ($N = 46$).

Algorithm	No. of studies	Studies
Artificial Neural Networks	20	Barbosa et al. (2020); Basir et al. (2021); Chen and Jing (2017); Choudhury et al. (2021); Danilevicz et al. (2021); Eugenio et al. (2020); Fieuzal et al. (2017); Garzia-Martinez et al. (2020); Habyarimana and Baloch (2021); Kross et al. (2020); Niedbala et al. (2019a; 2019b); Ozcan et al. (2021); Ramos et al. (2020); Sagan et al. (2021); Šestak et al. (2018); Sharifi (2021); Tian et al. (2021); Zhang et al. (2021); Zhou et al. (2021b)
Linear Regression	19	Adak et al. (2021); Barbosa et al. (2020); Basir et al. (2021); Barzin et al. (2020); Castaldi et al. (2015); Chen and Jing (2017); Choudhury et al. (2021); Fan et al. (2021); Feng et al. (2020); Florence et al. (2021); Habyarimana and Baloch (2021); Hassanzadeh et al. (2021); Meng et al. (2021); Ozcan et al. (2021); Ramos et al. (2020); Sagan et al. (2021); Šestak et al. (2018); Zhang et al. (2020); Zhou et al. (2021b)
Random Forests	18	Barbosa et al. (2020); Danilevicz et al. (2021); Fan et al. (2021); Fei et al. (2021a); Feng et al. (2020); Fieuzal et al. (2020); Filippi et al. (2019); Habyarimana and Baloch (2021); Hunt et al. (2019); Li et al. (2021); Meng et al. (2021); Ngie and Ahmed (2018); Ozcan et al. (2021); Ramos et al. (2020); Sagan et al. (2021); Wan et al. (2020); Wen et al. (2021); Zhou et al. (2021b)
Support Vector Regression	8	Barbosa et al. (2020); Choudhury et al. (2021); Fei et al. (2021a); Meng et al. (2021); Ramos et al. (2020); Sagan et al. (2021); Shafiee et al. (2021); Zhou et al. (2021b)
Gaussian Process Regression	6	Alebele et al. (2021); Choudhury et al. (2021); Fei et al. (2021a); Florence et al. (2021); Meng et al. (2021); Sharifi (2021)
Convolutional Neural Networks	5	Barbosa et al. (2020); Fajardo and Whelan (2021); Nevavuori et al. (2019); Nevavuori et al. (2020); Zhou et al. (2021a)
Gradient Boosting	4	Barzin et al. (2020); Danilevicz et al. (2021); Habyarimana and Baloch (2021); Zhang et al. (2021)
k-Nearest Neighbor	3	Meng et al. (2021); Ramos et al. (2020); Sharifi (2021)
LASSO	3	Adak et al. (2021); Shafiee et al. (2021); Zhang et al. (2021)
Decision Tree	3	Da Silva et al. (2020); Fajardo and Whelan (2021); Sharifi (2021)
Elastic Net Regression	2	Adak et al. (2021); Fei et al. (2021b)
Ensemble Learner	2	Fei et al. (2021a); Ramos et al. (2020)
Ridge Regression	2	Adak et al. (2021); Fei et al. (2021a)
Adaptive Boosting	1	Meng et al. (2021)
Bayesian Linear Regression	1	Alebele et al. (2021)
Bayesian Ridge Regression	1	Habyarimana and Baloch (2021)
Cubist Regression	1	Castaldi et al. (2015)
Sigcomp	1	Costa et al. (2022)

3.5. Model evaluation

We report the results for the evaluation phase divided into (1) the experimental manipulation of factors, (2) the assessment of performance using cross-validation and test sets, (3) performance metrics adopted, and (4) the relationship between prediction horizon and performance.

3.5.1. Experimental manipulation

Table 6 reveals that about half of the studies tested different ML algorithms and different prediction horizons. Eleven studies assessed the impact of two or more different vegetation indices on performance. Overall, the number of tested factors was either one (17), two (18), three (6), or four (3). Three studies did not manipulate any factor but administered a single experimental condition (da Silva et al., 2020; Eugenio et al., 2020; Zhou et al., 2021a).

Regarding ML algorithms, many studies compared the performance of linear regression modeling with at least one other algorithm that does not assume linear additive relationships between independent and dependent variables, such as ANN and RF. Thirteen of these fifteen studies provided evidence for the lower performance of the linear regression algorithm. With respect to different prediction horizons, 17 of 22 studies found better performance for smaller prediction horizons.

Table 6. Experimental manipulation in studies ($N = 46$).

Factor	No. of studies	Studies
ML algorithm	26	Adak et al. (2021); Alebele et al. (2021); Barbosa et al. (2020); Barzin et al. (2020); Basir et al. (2021); Castaldi et al. (2015); Chen and Jing (2017); Choudhury et al. (2021); Danilevicz et al. (2021); Fajardo and Whelan (2021); Fei et al. (2021a); Feng et al. (2020); Florence et al. (2021); Habyarimana and Baloch (2021); Meng et al. (2021); Nevavuori et al. (2020); Ozcan et al. (2021); Ramos et al. (2020); Sagan et al. (2021); Šestak et al. (2018); Shafiee et al. (2021); Sharifi (2021); Tian et al. (2021); Zhang et al. (2020; 2021); Zhou et al. (2021b)
Prediction horizon	22	Barzin et al. (2020); Castaldi et al. (2015); Fei et al. (2021a; 2021b); Feng et al. (2020); Fieuzal et al. (2017; 2020); Florence et al. (2021); Garcia-Martinez et al. (2020); Hassanzadeh et al. (2021); Hunt et al. (2019); Li et al. (2021); Nevavuori et al. (2019; 2020); Ngie and Ahmed (2018); Niedbala et al. (2019a; 2019b); Sagan et al. (2021); Shafiee et al. (2021); Sharifi (2021); Wan et al. (2020); Zhang et al. (2021)
Vegetation indices	11	Alebele et al. (2021); Choudhury et al. (2021); Fan et al. (2021); Fei et al. (2021b); Florence et al. (2021); Hassanzadeh et al. (2021); Hunt et al. (2019); Ngie and Ahmed (2018); Ramos et al. (2020); Zhang et al. (2021); Zhou et al. (2021b)
Feature combination	7	Chen and Jing (2017); Garcia-Martinez et al. (2020); Meng et al. (2021); Ozcan et al. (2021); Shafiee et al. (2021); Wan et al. (2020); Wen et al. (2021)
Image data	3	Danilevicz et al. (2021); Fajardo and Whelan (2021); Sagan et al. (2021)
Feature selection method	2	Choudhury et al. (2021); Hassanzadeh et al. (2021)
Fertilization	2	Danilevicz et al. (2021); Wan et al. (2020)
Genotype	2	Costa et al. (2022); Kross et al. (2020)
Number of examples	2	Filippi et al. (2019); Wan et al. (2020)
Agro-ecological zone	1	Zhang et al. (2021)
Canopy	1	Florence et al. (2021)
Irrigation	1	Fei et al. (2021a)
Planting date	1	Adak et al. (2021)

3.5.2. Model assessment

Table 7 presents the results of the model assessment criteria. Three groups of studies can be identified. The first group includes 13 studies that only applied cross-validation. The second group comprises 14 studies that only used a test set for model evaluation. The last group adopted cross-validation to select models from a set of alternative models and then evaluated the models on a test set (19 studies). With respect to cross-validation, 4 of the 32 studies defined the folds temporally and 28 studies used cross-sectional folds. Among the 33 studies using a test set, seven test sets included unknown observations from a different season (temporal test set). Information about the size of the test set was available from 19 articles, and the size ranged from only 8 (Chen and Jing, 2017) to more than 2000 examples (Hunt et al., 2019; Nevavuori et al., 2019).

3.5.3. Performance metrics

Table 8 shows the results for the adoption of performance metrics. The most frequent metric was the RMSE (38), which is defined as the root of the mean square error, and thus has the same unit as the yield variable (i.e., kg per ha). Eleven studies indicated the normalized RMSE (defined as the RMSE divided by either the mean or range of the observed yield; otherwise, the type of normalization was not specified). The second-most frequent metric was R^2 (32), which measures how much of the variance in the yield variable can be determined by the features included in the prediction model. It indicates how good the prediction model fits to the data (a value of 1 represents a perfect fit and 0 stands for no fit). Thirteen studies reported the MAE (same unit as the yield variable), and nine studies reported the MAPE, which is the mean of the ratio between the absolute error and observed yield (percentage).

Table 7. Model assessment in studies ($N = 46$).

Study	Cross-validation:		Test set:		
	No. of folds	Type	Percentage	Examples	Type
Adak et al. (2021)	10	CS	–	–	–
Alebele et al. (2021)	–	–	NR	NR	CS
Barbosa et al. (2020)	5	CS	20+20	NR	CS
Barzin et al. (2020)	5	CS	–	–	–
Basir et al. (2021)	10	CS	11	20	CS
Castaldi et al. (2015)	–	–	NR	NR	CS
Chen and Jing (2017)	28	CS	22	8	CS
Choudhury et al. (2021)	10	CS	15	NR	CS
Costa et al. (2022)	5	CS	–	–	–
Danilevicz et al. (2021)	5	CS	10	443 (*)	CS
da Silva et al. (2020)	–	–	NR	NR	T
Eugenio et al. (2020)	5	CS	30	9	CS
Fajardo and Whelan (2021)	–	–	9	9 to 14	CS
Fan et al. (2021)	–	–	30	NR	CS
Fei et al. (2021a)	10	CS	10	84 (*)	CS
Fei et al. (2021b)	10	CS	–	–	–
Feng et al. (2020)	10	T	–	–	–
Fieuzal et al. (2017)	3	CS	33	10	CS
Fieuzal et al. (2020)	10	CS	–	–	–
Filippi et al. (2019)	a) NR; b) 3	a) CS; b) T	–	–	–
Florence et al. (2021)	10	CS	50	NR	T
García-Martínez et al. (2020)	–	–	15+15	NR	CS
Habyarimana and Baloch (2021)	5	CS	30	NR	CS
Hassanzadeh et al. (2021)	5	CS	–	–	–
Hunt et al. (2019)	10	CS	30	2638 (*)	CS
Kross et al. (2020)	–	–	a) 50; b) NR	NR	a) CS; b) T
Li et al. (2021)	7	T	–	–	–
Meng et al. (2021)	5	CS	50 (*)	NR	T
Nevavuori et al. (2019)	3	CS	15	2280 (*)	CS
Nevavuori et al. (2020)	5	CS	30	NR	CS
Ngie and Ahmed (2018)	–	–	33	32, 36	CS
Niedbała et al. (2019a)	–	–	15+15	45+45	T
Niedbała et al. (2019b)	–	–	15+15	44+44	T
Ozcan et al. (2021)	10	CS	–	–	–
Ramos et al. (2020)	10	CS	–	–	–
Sagan et al. (2021)	–	–	30	64, 87 (*)	CS
Šestak et al. (2018)	NR	CS	50	18	CS
Shafiee et al. (2021)	10	CS	30	119	CS
Sharifi (2021)	NR	CS	–	–	–
Tian et al. (2021)	10	T	–	–	–
Wan et al. (2020)	–	–	NR	100	T
Wen et al. (2021)	10	CS	20	231	CS
Zhang et al. (2020)	–	–	40	106	CS
Zhang et al. (2021)	10	CS	30	NR	CS
Zhou et al. (2021a)	–	–	30	NR	CS
Zhou et al. (2021b)	5	CS	30	100	CS
Count	32	n/a	33	n/a	n/a

Note. CS = cross-sectional. NR = not reported. T = temporal. (*) = no exact figures reported.

Table 8. Performance metrics reported in studies ($N = 46$).

Study	RMSE	R^2	NRMSE	MAE	MAPE	r	Other
Adak et al. (2021)	Yes	Yes	–	Yes	–	Yes	–
Alebele et al. (2021)	Yes	Yes	–	–	–	–	–
Barbosa et al. (2020)	–	–	–	–	–	–	Yes
Barzin et al. (2020)	Yes	Yes	–	–	–	–	–
Basir et al. (2021)	Yes	Yes	–	–	–	–	Yes
Castaldi et al. (2015)	Yes	–	–	–	–	–	Yes
Chen and Jing (2017)	Yes	Yes	Yes (1)	–	–	–	–
Choudhury et al. (2021)	Yes	Yes	–	Yes	–	–	Yes
Costa et al. (2022)	Yes	–	–	–	Yes	–	Yes
Danilevicz et al. (2021)	Yes	Yes	Yes (2)	–	–	–	–
da Silva et al. (2020)	Yes	Yes	–	–	–	–	Yes
Eugenio et al. (2020)	Yes	–	–	Yes	–	Yes	Yes
Fajardo and Whelan (2021)	Yes	–	–	–	–	–	Yes
Fan et al. (2021)	Yes	Yes	–	–	–	–	–
Fei et al. (2021a)	Yes	Yes	–	–	–	–	–
Fei et al. (2021b)	Yes	Yes	Yes (1)	Yes	–	–	–
Feng et al. (2020)	Yes	–	–	–	Yes	Yes	Yes
Fieuzal et al. (2017)	Yes	Yes	–	–	–	–	–
Fieuzal et al. (2020)	Yes	Yes	Yes (3)	–	–	–	–
Filippi et al. (2019)	Yes	–	–	–	–	–	Yes
Florence et al. (2021)	Yes	Yes	–	–	Yes	–	–
García-Martínez et al. (2020)	Yes	–	–	Yes	–	Yes	–
Habyarimana and Baloch (2021)	Yes	Yes	–	Yes	Yes	–	Yes
Hassanzadeh et al. (2021)	Yes	Yes	Yes (3)	–	–	–	–
Hunt et al. (2019)	Yes	Yes	–	–	–	–	–
Kross et al. (2020)	–	–	–	–	–	Yes	Yes
Li et al. (2021)	–	–	Yes (1)	–	–	Yes	–
Meng et al. (2021)	Yes	Yes	–	–	–	–	–
Nevavuori et al. (2019)	–	–	–	Yes	Yes	–	–
Nevavuori et al. (2020)	–	Yes	–	Yes	Yes	–	–
Ngie and Ahmed (2018)	Yes	Yes	–	–	–	–	–
Niedbała et al. (2019a)	Yes	–	–	Yes	Yes	–	Yes
Niedbała et al. (2019b)	Yes	–	–	Yes	Yes	–	Yes
Ozcan et al. (2021)	Yes	Yes	–	–	–	–	–
Ramos et al. (2020)	–	–	–	Yes	–	Yes	–
Sagan et al. (2021)	–	Yes	Yes (1)	–	–	–	–
Šestak et al. (2018)	Yes	Yes	–	–	–	–	–
Shafiee et al. (2021)	–	Yes	–	–	–	–	Yes
Sharifi (2021)	Yes	Yes	–	Yes	–	–	–
Tian et al. (2021)	Yes	Yes	Yes (3)	–	Yes	–	–
Wan et al. (2020)	Yes	Yes	Yes (1)	–	–	–	–
Wen et al. (2021)	Yes	Yes	–	–	–	–	Yes
Zhang et al. (2020)	Yes	Yes	–	Yes	–	–	–
Zhang et al. (2021)	Yes	Yes	Yes (1)	–	–	–	–
Zhou et al. (2021a)	Yes	Yes	Yes	–	–	–	–
Zhou et al. (2021b)	Yes	Yes	–	–	–	–	–
Count	38	32	11	13	9	7	16

Note. MAE = mean absolute error. MAPE = mean absolute percentage error. NRMSE = normalized RMSE by (1) mean, (2) range, or (3) not specified. r = Pearson correlation coefficient. R^2 = coefficient of determination. RMSE = root mean square error.

3.5.4. Relationship between prediction horizon and performance

Twenty-three studies trained at least one prediction model for a specific horizon and reported R^2 (details on the data extraction are available in the Supplement). This set of studies allowed us to explore the relationship between prediction horizon and R^2. The scatter plot shown in Fig. 3 presents the results for 30 prediction models from these studies and indicates the respective grain crops. For predictions between 0.8 and 2.7 months before harvest, the majority of models yielded an R^2 of at least 0.81, and the best models had an R^2 of between 0.86 and 0.92. Performance was much lower for models that used larger prediction horizons, such as 0.567 for a horizon of four months and 0.495 for eight months.

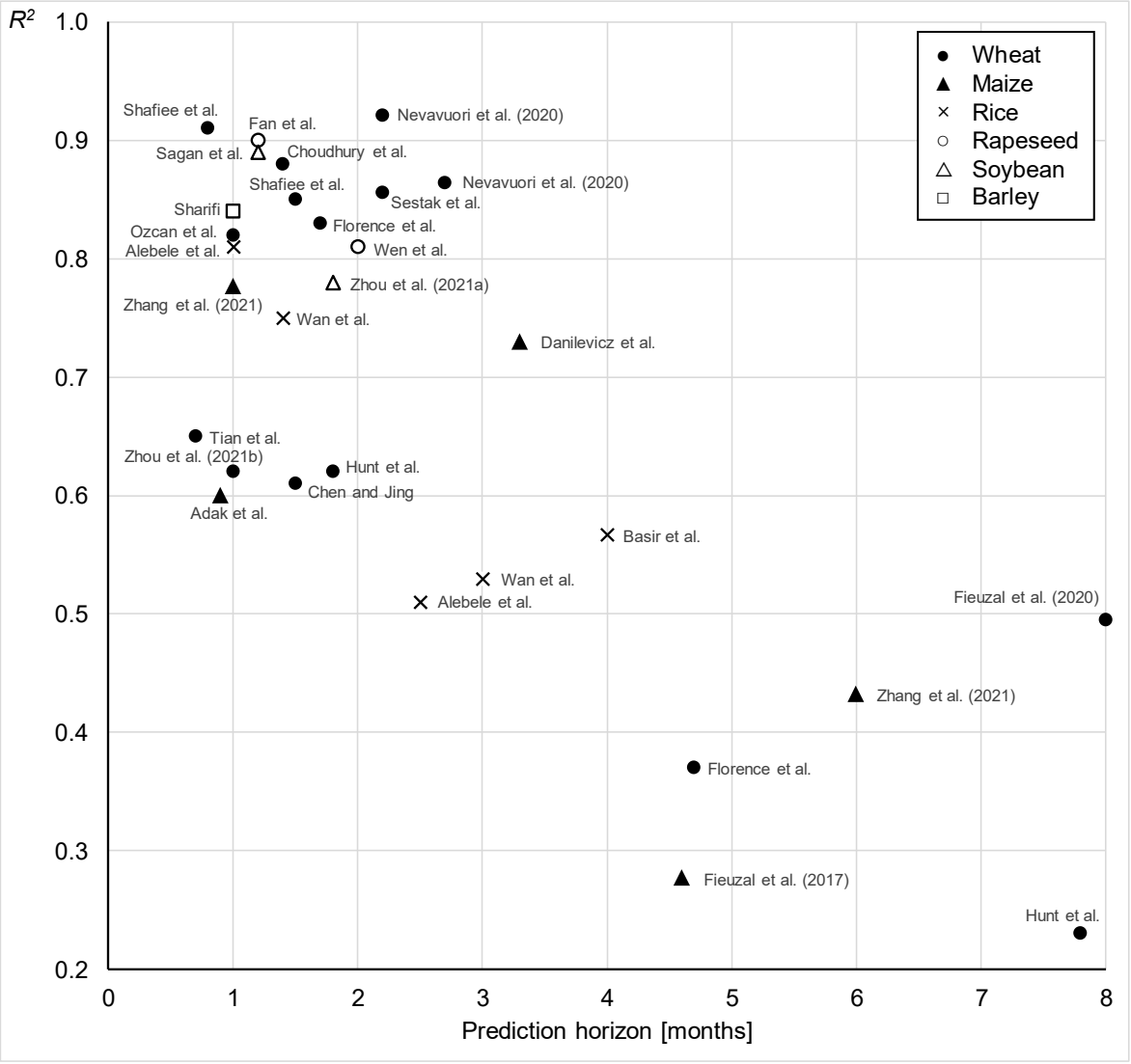

Fig. 3. R^2 by prediction horizon for 30 models.

Figure 4 shows the performance results based on the prediction time relative to the growth period (as a measure of how far along the growth period is). Six models made predictions when the season had progressed less than 30 percent. On the other hand, twelve models predicted

yields much later when more than 70 percent of the growth period had passed. For the former group of models, we note that the R^2 varied greatly between 0.23 and 0.92, while it was at least 0.60 for the latter group of models. Overall, variance declined with increasing progress of the growth period.

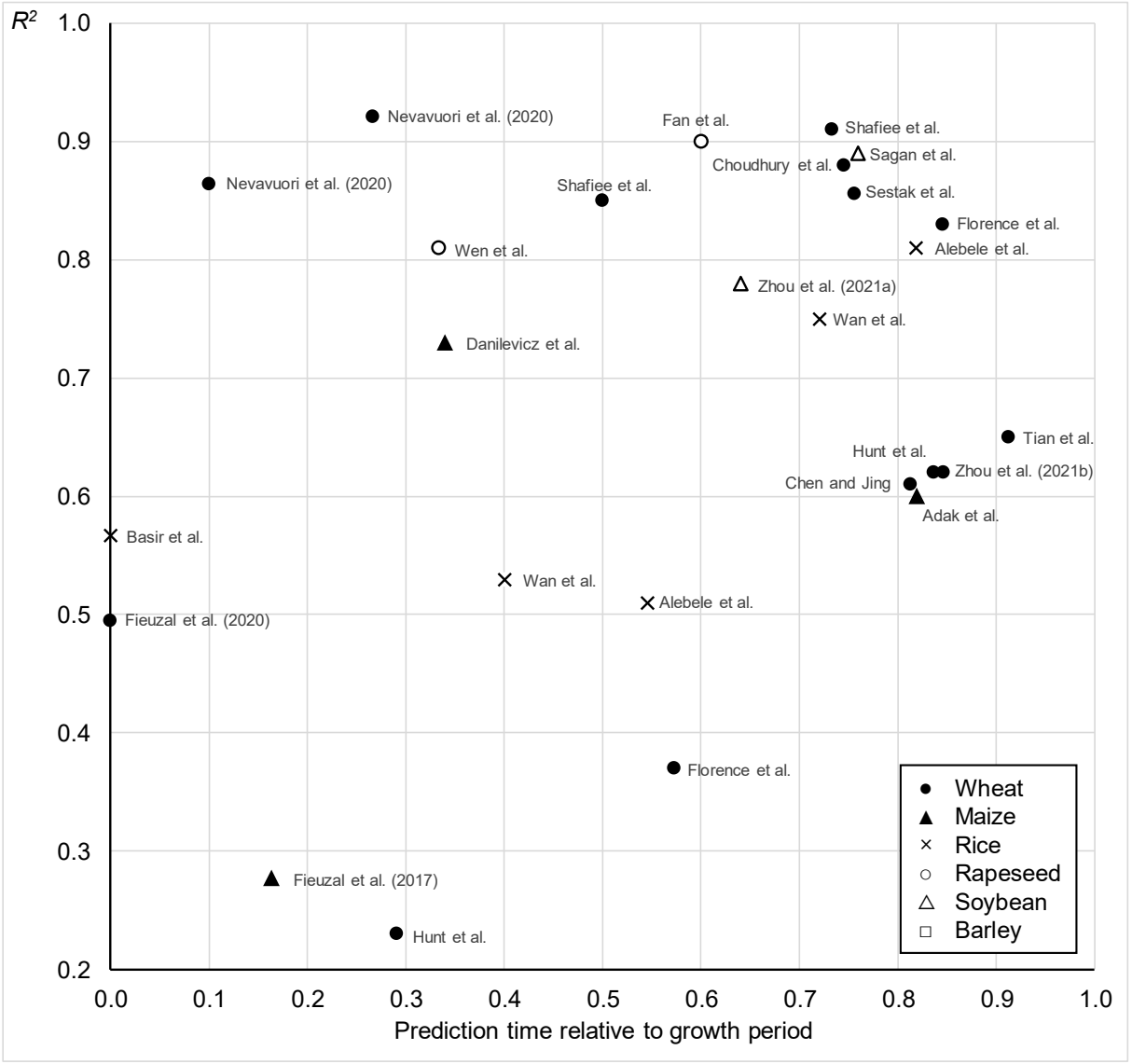

Fig. 4. R^2 by prediction time relative to growth period for 26 models.

4. Discussion

4.1. Principal findings and implications

This review analyzed forty-six studies that applied ML technology to predict field-scale grain yield in a season. The studies addressed yield prediction for the full range of vegetation stages of grain crops. Specifically, the prediction times ranged from the earliest date possible, i.e., shortly before or at the time of sowing, to a few weeks before harvest. Even the latter prediction times can assist crop farmers' decisions regarding corrective measures used to enhance yield (e.g., delaying the time of harvest). In addition, almost one-half of the studies tested different prediction horizons. These findings highlight the important role of the prediction horizon in the literature.

4.1.1. Data collection

With respect to the number of fields, seasons, and genotypes examined, no predominant type of study emerged. Collecting data for multiple genotypes grown on multiple fields in more than one season increases the number of examples available for model training and evaluation, thus bearing potential for the greater generalizability of a prediction model. However, each additional field and genotype requires extra effort for data collection, and adding one season doubles the time required for conducting a study. Therefore, it is not surprising that many studies focused on one genotype, field, or season. A noteworthy finding is the huge variety of field sizes (from hundreds of square meters to hundreds of hectare). This variety can partly be attributed to differences between studies conducted at research facilities and farms. The former studies enable better control of confounding factors, but this higher level of interval validity might decrease the variability of the data. Prediction models learned from that data might exhibit less generalizability compared to models learned from data collected in less controlled settings.

As in any predictive modeling scenario, a clear specification of the prediction problem and the data set used is necessary to be able to compare and integrate the results with related studies (Shmueli, 2010). While our review highlights the comprehensive coverage of prediction horizons, we also found considerable heterogeneity in the reporting. The studies adopted various units of measurements, such as time intervals, calendar dates, vegetation stages, and combinations of them. Some articles lacked sufficient information regarding the time when the last input data were recorded and the time of harvest, and many articles only provided rough details. Examples of good reporting are the study by Sagan et al. (2021), which provided the time of measurement for all input variables and the exact dates for sowing and harvesting, and the study by Fieuzal et al. (2020), which reported exact dates and visualized the performance results for different prediction horizons.

The data attributes considered in the studies comprised a broad coverage of factors associated with the growth of grain crops. On one hand, we identified studies that exclusively collected data for one category of attributes, such as weather (Nevavuori et al., 2020), management (Costa et al., 2022), soil (Fajardo and Whelan, 2021), and vegetation indices (Ramos et al., 2020). On the other hand, in every fifth study, the prediction model was trained with much richer data integrating weather, management, and vegetation indices. This difference is remarkable, given the

additional time and effort required to collect and preprocess these data originating from different sources. Another finding is the treatment of vegetation indices as the most frequently used category of input data: More than half of the studies used at least one VI but no weather information, assuming that all effects of weather on crop growth would be more or less reflected in the images from which the VIs were calculated. About one-third of the studies considered both types of attributes by using VIs as a direct measure of growth and weather as a causal factor of growth.

The high prevalence of vegetation indices in the studies mirrored in the types of sensors used, with half of the studies using data collected via satellites and other half using UAVs. The choice of sensor has major consequences for the development of a prediction model. Satellite images are readily available on a global scale for many decades; the acquisition and processing to calculate VIs require relatively low effort. However, this advantage comes at the cost of lower spatial resolutions, smaller frequencies, and ground images potentially covered by clouds. Our review show that the spatial resolution of images retrieved from satellites varies greatly (from 10 to 1000 meters per pixel). It is evident that the selection of satellite imagery should be aligned to the field size, so that reflectance information sufficiently captures the within-field variability. Against this backdrop, the level of reporting in that regard is rather low. UAVs produce images of much higher resolution, and the time of recording is determined by the developer's requirements, though each field survey requires considerable effort. A common theme in the literature is the empirical evaluation of satellites vis-à-vis UAVs as a basis for yield predictions, and our review corroborates this theme for the early prediction of grain yield. The results of single studies aid understand the usefulness of either type of sensor, but the appraisal of evidence is still hindered by the heterogeneity of studies.

4.1.2. Data preprocessing

Feature selection is an important phase in the development of prediction models because it allows for reductions in the often-high dimensionality of the used data set. In grain yield prediction, this dimensionality originates from the various categories of input data and the many possible attributes within each category. Therefore, we expected that a large percentage of studies applied feature selection methods to focus on attributes that helped achieve high performance and omitted attributes that represented noise in the data. This expectation was partly met, with less than half of the studies reporting on feature selection. These studies were indeed successful in reducing the number of features without losing prediction performance, and some studies showed that models trained with fewer features exhibited greater performance. For instance, Barzin et al. (2020) chose a three-feature model because performance decreased when more features were added, and Chen and Jing (2017) chose a six-feature model that was computed from 14 candidate features.

It is noteworthy that some studies performed an analysis of the features but did not take advantage of the findings to reduce their number. These studies determined the relative feature importance using in-built techniques of the respective ML algorithms. For instance, in a study by Wen et al. (2021), 5 of 17 features exhibited extremely low importance-indices (i.e., seed rate, preceding crop, fertilizer method, plant density, and soil pH). Therefore, it is likely that a more parsimonious model with less features would have performed equally well. Similarly, a study by

Zhou et al. (2021b) reported a very low importance for plant height but retained the feature in the model.

Collectively, the results of our review indicate that feature selection can be a useful technique to reduce the high dimensionality of data sets. The variety of methods used suggests that researchers are aware of the available techniques. We also note studies that proposed domain-specific approaches to feature selection and found evidence for their usefulness (Hassanzadeh et al., 2021; Shafiee et al., 2021). However, the evidence base regarding feature selection was rather small. Therefore, we recommend that future research integrate feature selection as a standard technique into the development of yield prediction models (provided the number of features is not small). We also suggest examining alternative feature selection methods through experimental evaluation. Feature selection appears specifically promising due to the large number of different but related vegetation indices that are available from remote sensing.

4.1.3. Model training

The most used ML algorithms were ANN, LR, and RF. This finding largely corroborates observations of previous reviews examining crop yield prediction in general (Bali and Singla, 2022; Benos et al., 2021; van Klompenburg et al., 2020). In more than three-fourths of the studies, at least one of the three algorithms was tested. This dominance was coincident with only a handful of studies using a deep learning (DL) algorithm. Specifically, Convolutional Neural Networks were present in five studies, whereas no other DL algorithm (such as Long Short-Term Memory Networks, Recurrent Neural Networks, and Multilayer Perceptrons) was included in any study. In view of the improvements that DL has enabled to speech recognition, object detection, and many other domains (LeCun et al., 2015), it would be interesting to know how useful DL algorithms are in predicting continuous grain yield variables compared with the more traditional ML algorithms.

It is unfortunate that about 45% of the articles provided no information on the size of the training set used. This finding is worrisome for two reasons. First, these studies cannot support developers' estimatation of the number of input-output pairs that are required or reasonable for the training of a prediction model. Second, if a prediction model is trained from too few example pairs, its mapping function will become unreliable, even though the model might exhibit relatively good performance in the evaluation. The robustness of the model might be limited such that the model will perform worse in a different evaluation. For readers to be able to assess the adequacy of a training set, they must at least know its size. To overcome the current situation, articles should clearly indicate the absolute and relative size of the training set, which can be complemented with the same information for the test set (if used). All this information can elegantly be summarized in one sentence, such as the following: "[..] the dataset was divided into train (70%, 420 samples) and test (30%, 119 samples) sets" (Shafiee et al., 2021, p. 4).

4.1.4. Model evaluation

The effectiveness of a trained prediction model must be rigorously assessed via well-executed evaluation techniques. Evaluation is a crucial phase because there can be no ML prediction model that is a-priori superior to another model (Wolpert, 1996). For instance, a model trained from input

data for one field might perform equally well, much better, or much worse when applied to new data from a different field, even if the two fields are similar. The burden is on the developer to demonstrate that the model does not overfit the training set. Overfitting characterizes a model that learned the examples in the training set too well such that its performance on new data is negatively impacted. The results of our review show that more than 70% of the studies conducted the evaluation on a separate test set of new data. This approach represents the "gold standard" in ML, whereas cross-validation is the preferred technique if the available data are too small to divide into sufficiently large training and test sets. Despite the importance of evaluation on new data, only 19 of 33 studies reported the size of the test set. Although larger test sets require more effort for data collection, they enable the better assessment of performance. Therefore, the size of the test set should be included in the reporting and the results should be discussed in view of that size.

With respect to the adoption of performance metrics, we believe that the findings of our review have four important implications for future research. First is the duality of metrics with units of measurement (e.g., RMSE and MAE) and metrics with no units (e.g., R^2 and NRMSE). The former metrics can be directly interpreted in the domain because their units are the same as the yield variable; hence, they are can be used to inform management decisions. The latter metrics enable collation of performance across different grains, fields, seasons, and data sets. With each group of metrics serving a different purpose, studies should include metrics of both groups in their reporting to address these purposes. This approach can help paint a more comprehensive picture of performance.

Second is the great variety of metrics, which makes the comparison and integration of evidence from individual studies difficult. The results of different studies can only be collated if the same metrics have been reported (although this condition is not sufficient). While the RMSE and R^2 were the most frequently used metrics, they only accounted for 80% and 70% of the studies, respectively. We recommend reporting a broader set of standard metrics in every study to extend the evidence base for meta-analysis.

Third is the incidence of metrics that are either less useful or inadequate, and should thus be replaced or abandoned from the reporting. One advantage of the MAE over the RMSE is that any difference between predicted and observed yield has a proportional effect on the metric. Although this advantage has been identified in other studies (Chai and Draxler, 2014; Willmott and Matsuura, 2005), only every fourth study reported the MAE. Similarly, the MAPE can be a useful replacement for the NRMSE, and the MAE and MAPE should then be reported in combination. Regarding unitless metrics, the Pearson correlation coefficient was present in seven studies, even though a linear relationship does not demonstrate that a prediction error is small (Sheiner and Beal, 1981). For instance, let the predicted yield always be 40% greater than the observed yield. In this case, the error is very large (MAPE = 0.4), whereas the correlation is perfect ($r = 1$). In other words, the correlation coefficient does not tell us anything about the magnitude of error. In addition, this coefficient assumes that grain yields are uniformly distributed, which is often not the case or studies did not report on whether the assumption was fulfilled.

Fourth is the much lower R^2 of predictions that were made many months before harvest, as signified by the sharp drop to less than 0.6 for predictions four or more months before harvest. On the other hand, most studies used smaller prediction horizons (up to 2.7 months), and the majority of these models achieved an r-squared of at least 0.81. The latter finding suggests that – contrary to our hypothesis – the relationship between prediction horizon and R^2 is not monotonically negative. Noting that few studies addressed predictions between three and eight months before harvest, we suggest that future research address this range to test the boundaries of very high performance. For instance, a recent study demonstrated how the most accurate predictions of sugarcane yield (which is not a grain crop) can be made one month earlier than in previous studies (Akbarian et al., 2022). In a similar vein, our analysis of the prediction time relative to the growth period showed that prediction models for predictions very early in the season are underrepresented and exhibit greater variance of R^2 than models for predictions late in the season.

Taken together, the heterogeneity in the reporting of performance metrics represents a barrier to the integration of evidence for the effectiveness of prediction models. This barrier hinders a more comprehensive synthesis of study results. Such synthesis can aid understanding of the relationships between factors considered in the design of yield prediction models and the facets of prediction performance. Our examination of the relationship between prediction horizon and R^2 (Section 3.5.4) uncovered that the prediction horizon plays an important but nuanced role. Therefore, conclusions drawn about the superiority of a specific ML technique (or prediction model) over another must take this relationship into account. The validity of conclusions can be enhanced by focusing on studies that tested similar prediction horizons and integrating prediction horizon as a confounding factor into the synthesis.

4.2. Limitations

Some limitations of our review must be noted. First, although we focused on early prediction at the field scale for grain crops, the set of studies exhibited considerable variance in the design of experiments, input data used, and assessment of prediction performance. Therefore, it was not possible to collate the performance results of all studies, except for the synthesis of results for prediction horizon and R^2 (this synthesis was limited to 23 of 46 studies). Second, subgroup analyses for different types of grains were not possible due to the small sample size. Third, our review was restricted to the information available from the articles, so it is possible that researchers have adopted additional techniques, conducted further analyses, and considered the obtained results in their development and final evaluation. Space constraints of journals might have forced researchers to focus on key issues and findings, thus hindering a more elaborate reporting of the ML process of data collection, data preprocessing, model training, and model evaluation. Fourth, some studies reported no exact quantitative results but provided only rough indications or charts, which made the data extraction intricate.

5. Conclusion

This systematic review examined the adoption of machine learning technology for the early prediction of grain yield at the field scale. The results provide insights into the richness of the ML techniques used for data collection, preprocessing, model training, and model evaluation. We identified five areas that bear potential to enhance the evidence base for the effectiveness of prediction models. First, we recommend further research to test the boundaries of very high performance, i.e., for prediction horizons greater than three months. The results of our quantitative synthesis suggest a non-monotonic relationship between prediction horizon and performance, and this finding was contrary to our expectation. Second, the large amount and variety of input data available for field-scale yield prediction appear attractive for the increased adoption of deep learning algorithms. Third, although field-level data often exhibit high dimensionality, in particular due to multiple vegetation indices, less than half of the studies applied feature selection. Fourth, a complete reporting of the number of examples included in the training and test sets can aid assessments of the robustness of the proposed models. Fifth, heterogeneity in the reporting of performance metrics is still a major barrier to the accumulation of evidence. This barrier can be easily overcome by reporting unitless metrics (i.e., R^2, NRMSE, and MAPE) along meaningful metrics with units (i.e., RMSE and MAE). Collectively, a greater uniformity of studies on grain yield prediction can facilitate the interpretation of individual studies, aid integration of the burgeoning results on alternative ML techniques, and ultimately better inform the development of accurate and robust prediction models.

Funding

This work was supported by the Federal Ministry of Food and Agriculture (BMEL) [grant: 28DE106A22], based on a decision of the Parliament of the Federal Republic of Germany, as part of the research project DiWenkLa (Digital Value Chains for a Sustainable Small-Scale Agriculture). DiWenkLa is also supported by the Ministry of Food, Rural Areas and Consumer Protection Baden-Württemberg.

CRediT author statement

Joerg Leukel: Conceptualization, Methodology, Investigation, Writing – Original Draft. **Tobias Zimpel:** Methodology, Investigation, Writing – Review & Editing. **Christoph Stumpe:** Methodology, Writing – Review & Editing.

Appendix A Supplementary Material

A supplement to this article can be found online at https://doi.org/10.1016/j.compag.2023.107721.

References

Adak, A., Murray, S.C., Božinović, S., Lindsey, R., Nakasagga, S., Chatterjee, S., Anderson, S.L., Wilde, S., 2021. Temporal vegetation indices and plant height from remotely sensed imagery can predict grain yield and flowering time breeding value in maize via machine learning regression. Remote Sens. 13, 2141. https://doi.org/10.3390/rs13112141.

Akbarian, S., Xu, C., Wang, W., Ginns, S., Lim, S., 2022. Sugarcane yields prediction at the row level using a novel cross-validation approach to multi-year multispectral images. Comput. Electron. Agric. 198, 107024. https://doi.org/10.1016/j.compag.2022.107024.

Alebele, Y., Wang, W., Yu, W., Zhang, X., Yao, X., Tian, Y., Zhu, Y., Cao, W., Cheng, T., 2021. Estimation of crop yield from combined optical and SAR imagery using Gaussian kernel regression. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 10520–10534. https://doi.org/10.1109/JSTARS.2021.3118707.

Bali, N., Singla, A., 2022. Emerging trends in machine learning to predict crop yield and study its influential factors: a survey. Arch. Computat. Methods Eng. 29, 95–112. https://doi.org/10.1007/s11831-021-09569-8.

Barbosa, A., Trevisan, R., Hovakimyan, N., Martin, N.F., 2020. Modeling yield response to crop management using convolutional neural networks. Comput. Electron. Agric. 170, 105197. https://doi.org/10.1016/j.compag.2019.105197.

Barriguinha, A., Castro Neto, M. de, Gil, A., 2021. Vineyard yield estimation, prediction, and forecasting: a systematic literature review. Agronomy 11, 1789. https://doi.org/10.3390/agronomy11091789.

Barzin, R., Pathak, R., Lotfi, H., Varco, J., Bora, G.C., 2020. Use of UAS multispectral imagery at different physiological stages for yield prediction and input resource optimization in corn. Remote Sens. 12, 2392. https://doi.org/10.3390/rs12152392.

Basir, M.S., Chowdhury, M., Islam, M.N., Ashik-E-Rabbani, M., 2021. Artificial neural network model in predicting yield of mechanically transplanted rice from transplanting parameters in Bangladesh. J. Agric. Food Inf. 5, 100186. https://doi.org/10.1016/j.jafr.2021.100186.

Basso, B., Liu, L., 2019. Seasonal crop yield forecast: methods, applications, and accuracies. Adv. Agron. 154, 201–255. https://doi.org/10.1016/bs.agron.2018.11.002.

Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D., Bochtis, D., 2021. Machine learning in agriculture: a comprehensive updated review. Sensors 21. https://doi.org/10.3390/s21113758.

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer, New York, 738 pp.

Castaldi, F., Casa, R., Pelosi, F., Yang, H., 2015. Influence of acquisition time and resolution on wheat yield estimation at the field scale from canopy biophysical variables retrieved from SPOT satellite data. Int. J. Remote Sens. 36, 2438–2459. https://doi.org/10.1080/01431161.2015.1041174.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Comput. Electr. Eng. 40, 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Chen, P., Jing, Q., 2017. A comparison of two adaptive multivariate analysis methods (PLSR and ANN) for winter wheat yield forecasting using Landsat-8 OLI images. Adv. Space Res. 59, 987–995. https://doi.org/10.1016/j.asr.2016.11.029.

Choudhury, R.M., Das, S., Christopher, J., Apan, A., Chapman, S., Menzies, N.W., Dang, Y.P., 2021. Improving biomass and grain yield prediction of wheat genotypes on sodic soil using integrated high-resolution multispectral, hyperspectral, 3D point cloud, and machine learning techniques. Remote Sens. 13, 3482. https://doi.org/10.3390/rs13173482.

Costa, L., McBreen, J., Ampatzidis, Y., Guo, J., Gahrooei, M.R., Babar, M.A., 2022. Using UAV-based hyperspectral imaging and functional regression to assist in predicting grain yield and related traits in wheat under heat-related stress environments for the purpose of stable yielding genotypes. Precision Agric. 23, 622–642. https://doi.org/10.1007/s11119-021-09852-5.

da Silva, E.E., Rojo Baio, F.H., Ribeiro Teodoro, L.P., da Silva Junior, C.A., Borges, R.S., Teodoro, P.E., 2020. UAV-multispectral and vegetation indices in soybean grain yield prediction based on in situ observation. Remote Sens. Appl.: Soc. Environ. 18, 100318. https://doi.org/10.1016/j.rsase.2020.100318.

Danilevicz, M.F., Bayer, P.E., Boussaid, F., Bennamoun, M., Edwards, D., 2021. Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection. Remote Sens. 13, 3976. https://doi.org/10.3390/rs13193976.

Eugenio, F.C., Grohs, M., Venancio, L.P., Schuh, M., Bottega, E.L., Ruoso, R., Schons, C., Mallmann, C.L., Badin, T.L., Fernandes, P., 2020. Estimation of soybean yield from machine learning techniques and multispectral RPAS imagery. Remote Sens. Appl.: Soc. Environ. 20, 100397. https://doi.org/10.1016/j.rsase.2020.100397.

Fajardo, M., Whelan, B.M., 2021. Within-farm wheat yield forecasting incorporating off-farm information. Precision Agric. 22, 569–585. https://doi.org/10.1007/s11119-020-09779-3.

Fan, H., Liu, S., Li, J., Li, L., Dang, L., Ren, T., Lu, J., 2021. Early prediction of the seed yield in winter oilseed rape based on the near-infrared reflectance of vegetation (NIRv). Comput. Electron. Agric. 186, 106166. https://doi.org/10.1016/j.compag.2021.106166.

Fei, S., Hassan, M.A., He, Z., Chen, Z., Shu, M., Wang, J., Li, C., Xiao, Y., 2021a. Assessment of ensemble learning to predict wheat grain yield based on UAV-multispectral reflectance. Remote Sens. 13, 2338. https://doi.org/10.3390/rs13122338.

Fei, S., Hassan, M.A., Ma, Y., Shu, M., Cheng, Q., Li, Z., Chen, Z., Xiao, Y., 2021b. Entropy weight ensemble framework for yield prediction of winter wheat under different water stress treatments using unmanned aerial vehicle-based multispectral and thermal data. Front. Plant Sci. 12, 730181. https://doi.org/10.3389/fpls.2021.730181.

Feng, P., Wang, B., Liu, D.L., Waters, C., Xiao, D., Shi, L., Yu, Q., 2020. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agric. For. Meteorol. 285-286, 107922. https://doi.org/10.1016/j.agrformet.2020.107922.

Fieuzal, R., Bustillo, V., Collado, D., Dedieu, G., 2020. Combined use of multi-temporal Landsat-8 and Sentinel-2 images for wheat yield estimates at the intra-plot spatial scale. Agronomy 10, 327. https://doi.org/10.3390/agronomy10030327.

Fieuzal, R., Marais Sicre, C., Baup, F., 2017. Estimation of corn yield using multi-temporal optical and radar satellite data and artificial neural networks. Int. J. Appl. Earth Obs. Geoinf. 57, 14–23. https://doi.org/10.1016/j.jag.2016.12.011.

Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D.S.N., Pozza, L.E., Ugbaje, S.U., Jephcott, T.G., Paterson, S.E., Whelan, B.M., Bishop, T.F.A., 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precision Agric. 20, 1015–1029. https://doi.org/10.1007/s11119-018-09628-4.

Florence, A., Revill, A., Hoad, S., Rees, R., Williams, M., 2021. The effect of antecedence on empirical model forecasts of crop yield from observations of canopy properties. Agriculture 11, 258. https://doi.org/10.3390/agriculture11030258.

García-Martínez, H., Flores-Magdaleno, H., Ascencio-Hernández, R., Khalil-Gardezi, A., Tijerina-Chávez, L., Mancilla-Villa, O.R., Vázquez-Peña, M.A., 2020. Corn grain yield estimation from vegetation indices, canopy cover, plant density, and a neural network using multispectral and RGB images scquired with unmanned aerial vehicles. Agriculture 10, 277. https://doi.org/10.3390/agriculture10070277.

Habyarimana, E., Baloch, F.S., 2021. Machine learning models based on remote and proximal sensing as potential methods for in-season biomass yields prediction in commercial sorghum fields. PLOS ONE 16, e0249136. https://doi.org/10.1371/journal.pone.0249136.

Hassanzadeh, A., Zhang, F., van Aardt, J., Murphy, S.P., Pethybridge, S.J., 2021. Broadacre crop yield estimation using imaging spectroscopy from unmanned aerial systems (UAS): a field-based case study with snap bean. Remote Sens. 13, 3241. https://doi.org/10.3390/rs13163241.

He, L., Fang, W., Zhao, G., Wu, Z., Fu, L., Li, R., Majeed, Y., Dhupia, J., 2022. Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods. Comput. Electron. Agric. 195, 106812. https://doi.org/10.1016/j.compag.2022.106812.

Ho, T.K., 1995. Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition (ICDAR). IEEE, pp. 278–282.

Hunt, M.L., Blackburn, G.A., Carrasco, L., Redhead, J.W., Rowland, C.S., 2019. High resolution wheat yield mapping using Sentinel-2. Remote Sens. Environ. 233, 111410. https://doi.org/10.1016/j.rse.2019.111410.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349, 255–260. https://doi.org/10.1126/science.aaa8415.

Kross, A., Znoj, E., Callegari, D., Kaur, G., Sunohara, M., Lapen, D.R., McNairn, H., 2020. Using artificial neural networks and remotely sensed data to evaluate the relative importance of variables for prediction of within-field corn and soybean yields. Remote Sens. 12, 2230. https://doi.org/10.3390/rs12142230.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Li, L., Wang, B., Feng, P., Wang, H., He, Q., Wang, Y., Liu, D.L., Li, Y., He, J., Feng, H., Yang, G., Yu, Q., 2021. Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. Agric. For. Meteorol. 308-309, 108558. https://doi.org/10.1016/j.agrformet.2021.108558.

López-Lozano, R., Duveiller, G., Seguini, L., Meroni, M., García-Condado, S., Hooker, J., Leo, O., Baruth, B., 2015. Towards regional grain yield forecasting with 1km-resolution EO biophysical products: strengths and limitations at pan-European level. Agric. For. Meteorol. 206, 12–32. https://doi.org/10.1016/j.agrformet.2015.02.021.

Meng, L., Liu, H., L. Ustin, S., Zhang, X., 2021. Predicting maize yield at the plot scale of different fertilizer systems by multi-source data and machine learning methods. Remote Sens. 13, 3760. https://doi.org/10.3390/rs13183760.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann. Intern. Med. 151, 264-9, W64. https://doi.org/10.7326/0003-4819-151-4-200908180-00135.

Morais, T.G., Teixeira, R.F., Figueiredo, M., Domingos, T., 2021. The use of machine learning methods to estimate aboveground biomass of grasslands: A review. Ecol. Indic. 130, 108081. https://doi.org/10.1016/j.ecolind.2021.108081.

Mouazen, A.M., Alexandridis, T., Buddenbaum, H., Cohen, Y., Moshou, D., Mulla, D., Nawar, S., Sudduth, K.A., 2020. Monitoring, in: Castrignano, A., Buttafuoco, G., Khosla, R., Mouazen, A., Moshou, D., Naud, O. (Eds.), Agricultural internet of things and decision support for precision smart farming. Elsevier, pp. 35–138.

Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N.H., Islam, N., 2022. A systematic literature review on crop yield prediction with deep learning and remote sensing. Remote Sens. 14, 1990. https://doi.org/10.3390/rs14091990.

Nevavuori, P., Narra, N., Linna, P., Lipping, T., 2020. Crop yield prediction using multitemporal UAV data and spatio-temporal deep Llarning models. Remote Sens. 12, 4000. https://doi.org/10.3390/rs12234000.

Nevavuori, P., Narra, N., Lipping, T., 2019. Crop yield prediction with deep convolutional neural networks. Comput. Electron. Agric. 163, 104859. https://doi.org/10.1016/j.compag.2019.104859.

Ngie, A., Ahmed, F., 2018. Estimation of maize grain yield using multispectral satellite data sets (SPOT 5) and the random forest algorithm. S. Afr. J. Geomat. 7, 11. https://doi.org/10.4314/sajg.v7i1.2.

Niedbała, G., Nowakowski, K., Rudowicz-Nawrocka, J., Piekutowska, M., Weres, J., Tomczak, R.J., Tyksiński, T., Álvarez Pinto, A., 2019a. Multicriteria prediction and simulation of winter wheat yield using extended qualitative and quantitative data based on artificial neural networks. Appl. Sci. 9, 2773. https://doi.org/10.3390/app9142773.

Niedbała, G., Piekutowska, M., Weres, J., Korzeniewicz, R., Witaszek, K., Adamski, M., Pilarski, K., Czechowska-Kosacka, A., Krysztofiak-Kaniewska, A., 2019b. Application of artificial neural networks for yield modeling of winter rapeseed based on combined quantitative and qualitative data. Agronomy 9, 781. https://doi.org/10.3390/agronomy9120781.

Oikonomidis, A., Catal, C., Kassahun, A., 2022. Deep learning for crop yield prediction: a systematic literature review. N. Z. J. Crop Hortic. Sci., 1–26. https://doi.org/10.1080/01140671.2022.2032213.

Olson, D., Anderson, J., 2021. Review on unmanned aerial vehicles, remote sensors, imagery processing, and their applications in agriculture. J. Agron. 113, 971–992. https://doi.org/10.1002/agj2.20595.

Ozcan, A., Leloglu, U.M., Suzen, M.L., 2021. Early wheat yield estimation at field-level by photosynthetic pigment unmixing using Landsat 8 image series. Geocarto Int., 1–17. https://doi.org/10.1080/10106049.2021.1903577.

Ramos, A.P.M., Osco, L.P., Furuya, D.E.G., Gonçalves, W.N., Santana, D.C., Teodoro, L.P.R., da Silva Junior, C.A., Capristo-Silva, G.F., Li, J., Baio, F.H.R., Marcato Junior, J., Teodoro, P.E.,

Pistori, H., 2020. A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices. Comput. Electron. Agric. 178, 105791. https://doi.org/10.1016/j.compag.2020.105791.

Raschka, S., 2015. Python machine learning. Packt Publishing.

Rashid, M., Bari, B.S., Yusup, Y., Kamaruddin, M.A., Khan, N., 2021. A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. IEEE Access 9, 63406–63439. https://doi.org/10.1109/ACCESS.2021.3075159.

Sagan, V., Maimaitijiang, M., Bhadra, S., Maimaitiyiming, M., Brown, D.R., Sidike, P., Fritschi, F.B., 2021. Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning. ISPRS J. Photogramm. Remote Sens. 174, 265–281. https://doi.org/10.1016/j.isprsjprs.2021.02.008.

Šestak, I., Mesić, M., Zgorelec, Ž., Perčin, A., 2018. Diffuse reflectance spectroscopy for field scale assessment of winter wheat yield. Environ. Earth Sci. 77. https://doi.org/10.1007/s12665-018-7686-x.

Shafiee, S., Lied, L.M., Burud, I., Dieseth, J.A., Alsheikh, M., Lillemo, M., 2021. Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. Comput. Electron. Agric. 183, 106036. https://doi.org/10.1016/j.compag.2021.106036.

Sharifi, A., 2021. Yield prediction with machine learning algorithms and satellite images. J. Sci. Food Agric. 101, 891–896. https://doi.org/10.1002/jsfa.10696.

Sheiner, L.B., Beal, S.L., 1981. Some suggestions for measuring predictive performance. J. Pharmacokinet. Biopharm. 9, 503–512. https://doi.org/10.1007/BF01060893.

Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M., Ebrahimie, E., 2014. Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture. PLOS ONE 9, e97288. https://doi.org/10.1371/journal.pone.0097288.

Shmueli, G., 2010. To explain or to predict? Statist. Sci. 25. https://doi.org/10.1214/10-STS330.

Tian, H., Wang, P., Tansey, K., Han, D., Zhang, J., Zhang, S., Li, H., 2021. A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the Guanzhong Plain, PR China. Int. J. Appl. Earth Obs. Geoinf. 102, 102375. https://doi.org/10.1016/j.jag.2021.102375.

van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: a systematic literature review. Comput. Electron. Agric. 177, 105709. https://doi.org/10.1016/j.compag.2020.105709.

Wan, L., Cen, H., Zhu, J., Zhang, J., Zhu, Y., Sun, D., Du, X., Zhai, L., Weng, H., Li, Y., Li, X., Bao, Y., Shou, J., He, Y., 2020. Grain yield prediction of rice using multi-temporal UAV-based RGB and multispectral images and model transfer – a case study of small farmlands in the South of China. Agric. For. Meteorol. 291, 108096. https://doi.org/10.1016/j.agrformet.2020.108096.

Wen, G., Ma, B.-L., Vanasse, A., Caldwell, C.D., Earl, H.J., Smith, D.L., 2021. Machine learning-based canola yield prediction for site-specific nitrogen recommendations. Nutr. Cycl. Agroecosyst. 121, 241–256. https://doi.org/10.1007/s10705-021-10170-5.

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82. https://doi.org/10.3354/cr030079.

Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.-J., 2017. Big data in smart farming – a review. Agric. Syst. 153, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023.

Wolpert, D.H., 1996. The lack of a priori distinctions between learning algorithms. Neural Comput. 8, 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341.

Xue, J., Su, B., 2017. Significant remote sensing vegetation indices: a review of developments and applications. J. Sens. 2017, 1–17. https://doi.org/10.1155/2017/1353691.

Zeng, L., Wardlow, B.D., Xiang, D., Hu, S., Li, D., 2020. A review of vegetation phenological metrics extraction using time-series, multispectral satellite data. Remote Sens. Environ. 237, 111511. https://doi.org/10.1016/j.rse.2019.111511.

Zhang, L., Zhang, Z., Luo, Y., Cao, J., Xie, R., Li, S., 2021. Integrating satellite-derived climatic and vegetation indices to predict smallholder maize yield using deep learning. Agric. For. Meteorol. 311, 108666. https://doi.org/10.1016/j.agrformet.2021.108666.

Zhang, P.-P., Zhou, X.-X., Wang, Z.-X., Mao, W., Li, W.-X., Yun, F., Guo, W.-S., Tan, C.-W., 2020. Using HJ-CCD image and PLS algorithm to estimate the yield of field-grown winter wheat. Sci. Rep. 10, 5173. https://doi.org/10.1038/s41598-020-62125-5.

Zhou, J., Zhou, J., Ye, H., Ali, M.L., Chen, P., Nguyen, H.T., 2021a. Yield estimation of soybean breeding lines under drought stress using unmanned aerial vehicle-based imagery and convolutional neural network. Biosyst. Eng. 204, 90–103. https://doi.org/10.1016/j.biosystemseng.2021.01.017.

Zhou, X., Kono, Y., Win, A., Matsui, T., Tanaka, T.S.T., 2021b. Predicting within-field variability in grain yield and protein content of winter wheat using UAV-based multispectral imagery and machine learning approaches. Plant Prod. Sci. 24, 137–151. https://doi.org/10.1080/1343943X.2020.1819165.