








# Wheat Yield Prediction Using Machine Learning: A Survey

Taye Girma Debelee<sup>1,2</sup> , Samuel Rahimeto Kebede<sup>1</sup> ,  
Fraol Gelana Waldamichael<sup>1</sup>  , and Daniel Moges Tadesse<sup>1</sup> 

<sup>1</sup> Ethiopian Artificial Intelligence Institute, 40782 Addis Ababa, Ethiopia  
tayegirma@gmail.com, taye.girma@aaii.et,

{samuel.rahimeto, fraol.gelana}@aic.et

<sup>2</sup> Addis Ababa Science and Technology University, College of Electrical  
and Mechanical Engineering, 120611 Addis Ababa, Ethiopia

**Abstract.** Wheat is one of the most important and most produced cereal crops in the world with over 600 million tonnes harvested annually. Accurate yield prediction of this important crop plays a huge role in the nation's plan for achieving sustainable food security. In this work, we performed a systematic review of research works conducted on the application of machine learning in wheat yield prediction. The reviewed papers are acquired from multiple digital libraries based on a defined article selection requirement and the primary research question we hope to answer. In total, we filtered 24 relevant research articles conducted between the years 2019 and 2022, and identified the state-of-the-art machine learning algorithms currently adopted and the types of datasets used. As such, we found that random forest and gradient boosting are efficient and reliable choices for the task of wheat yield prediction. We also observed the rising popularity of deep learning algorithms, such as deep convolutional neural networks and LSTMs for remote sensing and time series-based wheat yield prediction. We also identified the lack of a large public dataset as a major challenge as it makes the reproduction and comparison of different model performances very difficult.

**Keywords:** Yield Prediction · Crop · Wheat · Machine Learning · Deep Learning

## 1 Introduction

In the disciplines of academia, business, and particularly in healthcare for early detection, diagnosis, prediction, and classification [4, 5, 9–12], machine learning is used to address a number of difficulties. The development of efficient algorithms and reasonably priced yet powerful technology has made it viable to use machine learning and deep learning in the agriculture sector [1]. Machine learning and deep learning have several uses in the agricultural sector. Early diagnosis of plant

diseases has been successfully accomplished by machine learning [37], We have demonstrated this in our previous work [37] where we proposed an algorithm for detecting coffee leaf diseases using HSV color segmentation and deep learning that will enable farmers to monitor the health of their coffee farm using a smartphone. Similarly, Afework and Debelee [1] utilized deep learning for the detection of bacterial wilt on the Enset crop which is the main food for around 20 Million people in the southern part of Ethiopia.

Crop yield estimation is an essential factor in sustainable agriculture [21], and machine learning has been used to predict yields of various crops with success [27] of which one is the main focus of this review work. Wheat, being one of the most important crops being harvested with over 600 million tones of wheat harvested annually [29] is counted as among the “big three” kinds of cereal crops. The success of wheat is due to its adaptability and range of cultivation, growing in a wide range of geographic locations and weather conditions. Accurate yield prediction of this important crop is an immense economic and research interest. Much research around the use of machine learning in yield prediction of various crops has been done in recent years. Nevavuori et al. [24] performed crop yield prediction by using data from UAVs and deep learning. Here, the authors’ collected their data during the growing season by UAVs, the collected RGB image is then processed and fed into a deep learning algorithm to get the yield prediction. They showed that deep learning performed better than NDVI data and that the approach is suitable for predicting wheat and barley yield in a specific climate. Accurate prediction of yield patterns and identification of extreme yield loss causes in maize crops was successfully undertaken by Zhong et al. [41]. The authors developed a multi-task learning model that was used to achieve region-specific pattern recognition. The model takes in information about the environment and yields information and uses it to cluster the Corn Belt in the United States into several homogeneous regions. The proposed model then extracts temporal and soil patterns separately according to the specific input and network structure. In this work, we conducted a comprehensive systematic literature review of studies conducted on the application of machine learning in crop yield prediction especially focusing on wheat yield prediction; where we try to identify the most effective and state-of-the-art algorithms that are being applied and also the appropriate features required to enable an accurate yield prediction using the learning algorithms.

## 2 Related Works

There have been several reviews conducted on general crop yield projections. Finding review articles done exclusively for wheat yield prediction, on the other hand, proved problematic. As a result, we employed reviews of crop production projections using wheat as one of the recognized crop kinds. Oikonomidis et. al.

[25] performed a systematic literature review on predicting crop yield using deep learning techniques. They have identified 44 papers, of which only eight articles focus on wheat yield prediction. They also noted that CNN's are most used architecture for yield prediction.

Bali and Singla [3] explored various machine-learning techniques used in crop yield prediction and discussed the efficiency of hybrid models formed by combining multiple machine-learning techniques. The authors discussed the two crop yield estimation approaches, namely the crop growth model and the data-driven model. The authors implied that the mathematical crop growth models are efficient and can yield good results in the yield prediction of specific crops, but noted that these models are expensive to develop and are impractical for large-scale agricultural planning. On the contrary, the authors discussed that data-driven models are cheaper to develop and easier to deploy.

Klompenburg et al. [35] conducted a survey of machine learning techniques and features that are used in crop yield prediction. The authors reviewed 50 studies conducted on crop yield prediction using machine learning techniques and identified the most used features and algorithms. According to their study, the authors identified temperature, rainfall, and soil type as the most used features. Additionally, they identified convolutional neural networks as the most applied learning algorithm, followed by LSTM'S.

Muruganatham et al. [23] conducted a systematic review on the fusion of remote sensing and deep learning for the application of crop yield prediction. The review study was motivated by the desire to examine the influence of vegetation indices and discover how environmental conditions affect agricultural productivity. The authors set out to find the most regularly used features and deep learning architectures, and discovered that vegetation indices and meteorological data are the most commonly used features, while CNN and LSTM-based models are the most commonly used deep learning architectures.

Table 1. Summary of related works.

Authors	Contribution	Limitation
Oikonomidis et al. [25]	The review work concisely summarizes identified research works from different search engines on deep learning-based crop yield predictions	Since its focuses are the general crop yield prediction and it doesn't include a summary of each reviewed research work, determining which algorithms or datasets were suitable for wheat yield prediction is indeterminate
Bali and Singla [3]	The authors performed in detail discussion on the various crop yield estimation techniques, on the various factors affecting yield estimation and held a broad discussion on the deep learning in crop yield prediction	The work lacks a proper summary of the reviews, which makes it hard for the reader to grasp what kind of machine learning or deep learning approaches are effective and what type of data is mostly used in crop yield estimation
Klompenburg et al. [35]	Feature diagram enables the researchers to know the major features used in crop yield prediction, and the way the discussion section is organized	Each paper is not discussed well, instead, the paper explained different machine learning approaches
Muruganatham et al. [23]	The paper covered most deep learning-based wheat yield prediction using remote sensing data	The paper doesn't give an idea of how the yield prediction models and remote sensing data are efficient among different crop types

### 3 Methods

This systematic review (SLR) [20] work is intended to highlight new works on wheat yield prediction using machine learning approaches, including both classic machine learning algorithms and deep learning methods. The SLR stresses the need of having a well defined methodology for creating research questions, search methodologies for discovering relevant literature, and establishing the required exclusion and inclusion criteria for selecting the appropriate studies.

#### 3.1 Research Questions

In this study, we want to pinpoint machine learning methods used for crop yield prediction, particularly in the previous four years. Thus, the main research topic that we hope to address is:

PRQ: *“What cutting-edge machine learning methods have been employed in the last four years to forecast wheat yields?”*

In order to further assist in focusing the intended response to the core research question, secondary research questions are also prepared. These are:

- SRQ1: *What was the key motivation for applying machine learning for wheat yield prediction?*
- SRQ2: *What categories of data are utilized and accessible?*
- SRQ3: *Which key evaluation metrics are used to measure yield prediction?*
- SRQ4: *Which machine learning algorithm and dataset performed better for wheat yield prediction?*

#### 3.2 Search Strategies

We need to identify the right search strategies [20] in order to identify as many pertinent primary studies as possible that attempt to respond to the primary research question posed. We have defined our search strategy as follows:

- Choose different search databases for the recent publications related to the title.
- Decompose research questions for better search output.
- Create the keywords related to the title [20]
- Build **search strings** using “AND” and “OR” boolean.

The approach used to search for the primary studies was focused on five known search databases that include: Springer Link<sup>1</sup>, Science direct<sup>2</sup>, Wiley online library<sup>3</sup> and IEEE Xplore<sup>4</sup>. These databases were selected because they contain most machine learning-related papers.

To get the most out of the databases, an optimized and simplified search string need to be defined as indicated in Algorithm 1. Further, additional inclusion and exclusion criteria were defined as presented in Table 2.

<sup>1</sup> <https://link.springer.com/>.

<sup>2</sup> <https://sciencedirect.com/>.

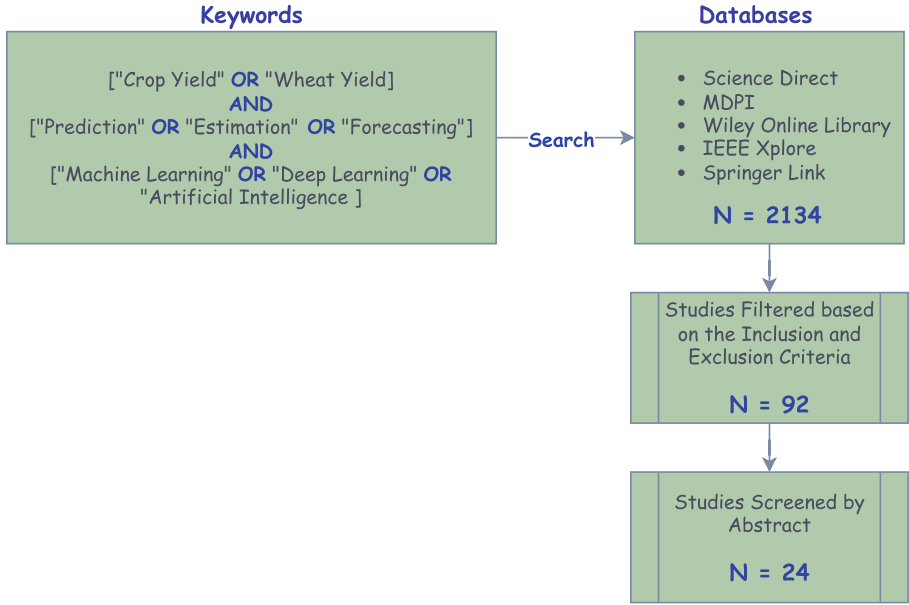
<sup>3</sup> <https://onlinelibrary.wiley.com/>.

<sup>4</sup> <https://ieeexplore.ieee.org/>.

**Algorithm 1.** Pseudo-code for defining search string.

```

Search.String = ("Crop Yield" OR "Wheat yield")
                AND
                ("Prediction" OR "Estimation" OR "Forecasting" )
                AND
                ("Machine Learning" OR "Deep Learning" OR "Artificial Intelligence")
    
```



**Fig. 1.** Methodology for systematic literature review.

**Table 2.** Paper selection criteria.

Inclusion criteria (IC)	Exclusion criteria (EC)
<b>IC1:</b> Studies that focused on wheat yield prediction	<b>EC1:</b> Duplicate publications
<b>IC2:</b> Studies carried out from 2019 to 2022.	<b>EC2:</b> Studies performed other than the English language
<b>IC3:</b> The article should be in reputable journals or recognized Conference proceedings.	<b>EC3:</b> MSc and Ph.D. thesis, Posters, Seminar, and Case studies
<b>IC4:</b> Publishing Journals should be indexed in web of science or Scopus	<b>EC4:</b> Studies that do not use either machine learning or deep learning

## 4 Machine Learning in Wheat Yield Prediction

### 4.1 Remote Sensing Based Wheat Yield Prediction

In order to estimate wheat yield in China, Zhou et al. [42] investigated the potential of nine climate factors, three metrics obtained from remote sensing, and three machine learning techniques. They discovered that the northern winter and spring wheat planting zones had the best results. Climate variables connected to water performed better than those related to temperature. In terms of predicting crop yield, they also found that solar-induced chlorophyll fluorescence outperformed the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI). The prediction in winter wheat planting zones performed better than the prediction in spring wheat planting zones, and the support vector machine outperformed other models.

In another study made by Kamir et al. [19] performed on the accurate estimation of yields from wheat across the Australian wheat belt using machine learning, and regression models were found to be more accurate than benchmark approaches and they were able to explain a significant amount of the yield variability observed across statistical units. The authors used data on yield, satellite images, and climate data. The satellite images were from the MODIS “MOD13Q1” data set, and the climate data from SILO (Scientific Information for Land Owners). The authors found the SVM algorithm which explained 77% of the wheat yield variability to be the best performing of the regression models and Ens.BDF to be the best ensemble model which explained 76% of the wheat yield variability with an RMSE of 0.57.

Tian et al. [34] built a model using an LSTM neural network technique and data from remote sensors and meteorology to improve wheat yield estimation. To estimate wheat production in the Guanzhong Plain, vegetation temperature condition index, climatic data, and leaf area index are very important, especially at the growth stages of wheat.

The model used several time steps to capture time series data with LSTM. The results showed that the best yield estimation accuracy (RMSE = 357.77 kg/ha and  $R^2 = 0.83$ ) was achieved with two-time steps and the input combination of meteorological data and two remote sensing indices. The authors compared the best LSTM model performance with BPNN and SVM for yield estimation accuracy. The LSTM model outperformed BPNN ( $R^2 = 0.42$  and RMSE = 812.83 kg/ha) and SVM ( $R^2 = 0.41$  and RMSE = 867.70 kg/ha) because of its recurrent neural network structure that can handle nonlinear relationships between multi-features inputs and yield. The authors also tested the optimal LSTM method on irrigation sites and rain-fed sites from 2008 to 2016 to check its robustness. The results showed that the proposed model was effective for different types of sampling sites and adaptable to inter-annual variations of climate.

Tesfaye et al. [33] undertook a study with the goal of developing a technique for remote sensing-based wheat production prediction in smallholding and heterogeneous agricultural settings. The study used vegetation indices from

high-resolution optical and SAR sensors to derive predictions. Five SAR indices were derived from the data of the S1 sensor, and eight vegetation indices from the S2 optical sensors. Data mining techniques were used, which fall into three major categories: statistical, machine learning, and deep learning. These techniques were used because of the intricate interaction between the predictors and response variables. Due to the limited availability of the response variable (field-collected wheat grain yield), this study used data mining techniques instead of the more typical approaches to machine learning and deep learning implementation. According to the study, networks with one or two hidden layers fared worse than deep neural networks with three hidden layers. The best models discovered using the three data mining techniques make use of phenological data, particularly data from the post-grain-filling period.

Vanli et al. [36] proposed a method of using satellite images to predict wheat yields in southeastern Turkey. The study found that the satellite images were accurate in predicting wheat yields, with an error of less than 200 kg/ha. In order to employ the optimum model for the geographical distribution of wheat crops, a total of eight machine-learning algorithms were evaluated and tuned for the categorization of satellite images. The machine learning algorithms produced outcomes with an accuracy of more than 90%. The random forest was chosen for picture categorization as the best model. With a root mean square error (RMSE) of 198 kg/ha, the tested model's observed and anticipated yields were relatively near to one another.

Fei et al. [15] applied five ML algorithms for fusing data from multiple sensors to predict crop yield more accurately. The ML algorithms were Cubist, SVM, DNN, RR, and RF. They used them for multi-sensor data fusion and ensemble learning for wheat grain yield prediction. The study showed that multi-sensor data was better than single-sensor data for prediction accuracy. The ensemble learning predictions had  $R^2$  values up to 0.692, which was higher than individual ML models with multi-sensor data. The RMSE, RPD, and RPIQ were 9160 kg/ha, 1.771, and 2.602, respectively. Their results indicated that low-altitude UAV-based multi-sensor data can be used for early grain yield prediction with data fusion and ensemble learning.

Another study in [14] examined the two machine learning methods of Bootstrapped Regression Trees (BRR) and Convolutional Neural Networks (CNN) and examined how they may be applied for predicting wheat yield. According to the study, local electromagnetic induction surveys or gamma radiometric surveys combined with BRR modeling utilizing publically available Sentinel data gave the most accurate estimates. With the addition of openly accessible data from related disciplines, the CNN models' outcomes improved.

Yang et al. [39] use the CERES wheat model to generate training samples for the training of their random forest model. Using the CERES wheat model simulations, they identified the leaf area index (LAI) and leaf nitrogen content (LNC) as the most sensitive parameters. These features were extracted from UAV's hyperspectral images and used as input into the CW-RF model to estimate winter wheat yield. The model (CERES wheat model) is not accurate



enough to be used as a ground truth for training the CW-RF model and needs further improvements.

Shidnal et al. [30] used machine learning to analyze how nutrient levels affect crop yield. They trained a neural network with Tensor Flow to recognize images of crops with different nutrient deficiencies (nitrogen, potassium, phosphorous) or healthy ones. They also used a clustering algorithm to measure the severity of the deficiency. Then they used a rule matrix to estimate the yield of the cropland based on the deficiency level. Their method was 76–77% accurate in predicting the yield.

Qiao et al. [28] developed a method for estimating crop yield using multi-spectral images. The method uses a 3-D convolutional neural network to extract features from the images that capture spatial and spectral information. Then, it uses a multi-kernel learning technique to combine the features from different domains. Finally, it uses a kernel-based method to get the probability distribution of the yield estimates. The method is tested on wheat yield prediction in China and compared with other methods. The results show that the method has  $R^2$  and RMSE values of 0.8 and 730 kg/ha respectively.

**Table 3.** Summary of papers on remote sensing-based wheat yield prediction.

Author	Method	Dataset	Acc	$R^2$	RMSE
Zhou et al. [42]	SVM	Remote sensing and climate data	-	0.63–0.74	1100 kg kg/ha
Kamir et al. [19]	SVM	Satellite and climate data	-	0.77	550 kg/ha
Tian et al. [34]	LSTM	Meteorological and remote sensing data	-	0.42	812.83 kg/ha
Tesfaye et al. [33]	DNN	Optical and radar data	-	-	1360 kg kg/ha
Vanli et al. [36]	Random Forest	Satellite image	90%	-	198 kg/ha
Fei et a. [15]	Ensemble learning	Multi-sensor data	-	0.692	916 kg/ha
Fajardo et al. [14]	Bootstrapped Regression Trees (BRR) and CNN	publicly available Sentinel data with the addition of local electromagnetic induction surveys or gamma radiometric surveys	-	-	600 kg/ha
Yang et al. [39]	Random Forest	UAV’s Hyperspectral Imagery & Synthetic Data from CERES wheat model	-	-	1,008.08 kg/ha
Shidnal et al. [30]	k-means clustering	Crop images	76–77%	-	-
Qiao et al. [28]	3-D convolutional neural network	Multispectral images	-	0.8	730 kg/ha

## 4.2 Environmental Factors Based Wheat Yield Prediction

Zhang et al. [40] proposed a generative adversarial networks (GANs) approach for increasing the precision of winter wheat yield estimation. GANs were proposed by the authors to deal with small datasets and a limited number of annotated samples. The training set consists of data from 2012 to 2015, while the

validation set consists of data from 2016. The test set is made up of data from 2017. GANs are used to supplement the training and validation sets. The CNN includes VTCI, LAI, and meteorological data. The CNN is better at predicting yield than previous models, according to the authors, because it can account for the interplay between several sorts of input characteristics.

Chergui [7] conducted durum wheat yield forecasting using machine learning and data augmentation to improve predictions and results. The author employed a dataset containing data on annual yields and acreage for harvest seasons ranging from 1991 to 2019. The study discovered that the data augmentation approach improved overall performance, with the Deep Neural Network producing the best results.

Pang et al. [26] proposed regional and local-scale wheat yield prediction using random forest regression (RFR). The Bureau of Meteorology provided data for this study, and collaborating farmers provided yield for the year 2018. The random forest regression technique was found to be accurate, with a high  $R^2$  value of 0.86 and a low RMSE of 0.18. The study also discovered that the technique was robust and worked well across a variety of paddocks with varying conditions.

Han et al. [18] investigated the use of several data sources and machine learning algorithms to estimate the winter wheat output in China, one or two months before harvest, they discovered that county-level models can forecast yield with good accuracy ( $R^2 > 0.7$  and error  $< 10\%$ ). They discovered that training intervals and agricultural zones have an impact on prediction accuracy. They made use of GEE and ArcGIS-processed remote sensing data. For the purpose of predicting yield, they examined three machine learning models (RF, GPR, and SVM). They claimed that RF outperformed GPR in terms of computation speed and accuracy.

Wang et al. [38] developed a method for estimating winter wheat yield within-season using various data sources in the US. The method tries to address the drawbacks of empirical models based on satellite images by using machine learning and multi-source data. The authors tested four machine learning models (SVM, RF, AdaBoost, and DNN) and reported that AdaBoost was the best. They also reported that decreasing the input factors enhanced the neural network's performance by preventing over-fitting and improving generalization ability.

The ABSOLUT v1.2 algorithm, which is used to forecast agricultural yields, was put forth by Tobias Conradt in [8]. The program uses correlations between time-aggregated meteorological indicators and agricultural yields to produce predictions. The method is used in Germany to predict the yields of important crops including winter wheat and silage maize. Separate training and testing years should be used when choosing features because the algorithm can make out-of-sample predictions (based only on data other than the target year to forecast).

In order to anticipate crop yields, Cao et al. [6] developed and used a hybrid skillful ML-dynamical model that blends ML with a global dynamical atmospheric prediction system. In their research, they examined multiple linear

regression (MLR) models as well as XGBoost, RF, and SVR. For the period of 2005 to 2014, they projected the production of winter wheat using three datasets: satellite data from MOD13C1, observational climate data from CRU, and S2S atmospheric prediction data from IAP CAS. With the S2S prediction as inputs, XGBoost outperformed the other four evaluated models, scoring  $R^2$  of 0.85 and RMSE of 780 kg/ha within 3–4 months before the winter wheat harvest. Their findings demonstrate that S2S dynamical forecasts outperform observational climate data for agricultural yield forecasting. Furthermore, their findings showed that integrating ML and S2S dynamical atmospheric prediction would be an advantageous yield forecasting tool, which might direct agricultural practices, policy, and agricultural insurance.

Murakami et al. [22] investigated meteorological limitations on winter wheat yield in Hokkaido, Japan's northernmost island, and compared ML models to a null model that returns the municipalities average yield to, neural network (NN), random forest (RF), support vector machine regression (SVR), partial least squares regression (PLS), cubist regression (CB), and multiple linear regression model (MLR). This island has a wet climate due to higher annual precipitation and an abundant snow-melt water supply in spring when compared to other wheat-producing areas. Their research discovered that precipitation, daily minimum air temperature, and irradiance had major effects on yield across the island during the grain-filling period. The study used 10-day mean meteorological data from seeding to harvest as predictor variables, as well as a one-year leave-out cross-validation procedure. The PLS, SVR, and RF had root means square errors of 872, 982, and 1,024 kg/ha, respectively, which were less than MLR (1,068 kg/ha) and the null model (1,035 kg/ha). Other metrics, such as Pearson's correlation coefficient and Nash-Sutcliffe efficiency, showed that these models outperformed the controls. The findings corroborated the authors' understanding of meteorological effects on wheat yield, implying the utility of explainable machine learning in meteorological crop yield prediction in wet climates.

Elavarasan and Vincent [13] proposed a reinforced random forest model for improved crop yield prediction by integrating agrarian parameters. The study describes a new algorithm developed to predict crop yield based on climate, soil, and water parameters. The Reinforcement Random Forest algorithm is a hybrid of regression and machine learning. Because it employs reinforcement learning, this new algorithm is expected to outperform other traditional machine learning techniques. This means that the algorithm learns from its errors and improves over time. The algorithm is also said to perform better with sparse data structures. The results showed that the proposed approach performs better, with lower error measures and an improved accuracy of 92.2%.

Using machine learning and multilayered, multifarm data sets, Filippi et al. developed a method to predict grain crop yield [17]. The authors outlined how crop yield models may be created using machine learning using data from various fields, farms, and years. In a case study, they used yield data from three seasons (2013-2015) spanning hundreds of hectares on substantial farms in Western Australia. For modeling, the yield data were cleaned up and combined into a grid of 100 m. Based on pre-sowing, mid-season, and late-season circumstances, they

projected wheat, barley, and canola yields using random forest models. They discovered that as additional within-season data became available, the models' accuracy increased (e.g. rainfall).

Ali et al. [2] suggested a two-phase universal ML model for predicting wheat yield. The model was based on online sequential extreme learning machines and ant colony optimization, and it utilised data from 27 counties in the agroecological zone. The ACO-OSELM model projected future yield at six test sites using yield data from a prior year as an input. Using a feature selection technique, ACO assisted in locating suitable data stations for the model's training and testing. In regions where historical crop data was substantially correlated, the hybrid ACO-OSELM model proved beneficial as a system for predicting crop yield.

**Table 4.** Summary of papers on environmental factors based wheat yield prediction.

Author	Method	Dataset	Acc	$R^2$	RMSE
Zhang et al. [40]	CNN with GAN	Environmental and remote sensing data	–	0.5	591.46 kg/ha
Nabila Chergui [7]	DNN	Historical yield data and climate data	–	0.96	4 kg/ha
Pang et al. [26]	Random forest regression	Meteorological data	–	0.45	250 kg/ha
Han et al. [18]	Random Forest	Soil data	–	0.75	6.89 kg/ha
Wang et al. [38]	AdaBoost	Historic yield records, remote sensing images, climate data, and soil maps.	–	–	510 kg/ha
Conrad. [8]	ABSOLUT v1.2 algorithm	Temperature, precipitation, and sunshine duration weather variables that are aggregated over different seasonal periods preceding the harvest	87.8%	–	115 kg/ha
Cao et al. [6]	MHCF v1.0	MOD13C1 satellite data, 225 CRU observational climate data, and IAP CAS S2S atmospheric prediction data	–	–	780 kg/ha
Murakami et al. [22]	partial least squares regression model	meteorological data	–	0.76	872kg/ha
Elavarasan and Vincent [13]	Reinforcement Random Fores	climate, soil and water data	92.2%	0.87	230kg/ha
Filippi et al. [17]	Random Forest	Multi fields, multi-farm and multi-seasonal data	–	0.85 to 0.92	0.36 to 420 kg/ha
Ali et al. [2]	Online sequential extreme learning machines coupled with ant colony optimization (ACO-OSELM)	27 agricultural counties' data within the Agro-ecological zone	–	0.968	155.86 kg/ha

### 4.3 Genomic and Phenology Based Wheat Yield Prediction

Feng et al. [16] present a mix of machine learning and bio-physical modeling to solve the typical constraints of frequently utilized statistical approaches for seasonal yield forecasting. This author investigated two methods for predicting wheat yields, multiple layer regression (MLR) and random forest (RF) models, and discovered that the RF model predicted observed yields better than the MLR model, especially in years with atypical yields, and provided better forecasts at earlier growth stages. The research was carried out in the New South Wales wheat area, which is located in southern Australia. The scientists integrated a crop simulation model with a statistical regression-based model in this work to dynamically anticipate wheat production at various stages throughout the growing season. APSIM is the crop simulation model, whereas RF is the regression-based model. The authors discovered that their hybrid model, which takes use of the strengths of each model, produced good yield prediction results. This was due mostly to the hybrid model's ability to utilize biophysical processes among crop, soil, management, and climate information, as well as a machine learning approach to account for climatic extremes and remote sensing data. Also, the machine learning technology utilized in the study outperformed standard regression methods.

Using genetic markers, genomic prediction (GP) is a technique for figuring out complex phenotypes. Increased grain production is essential for feeding the world, especially for basic crops like rice and wheat. Recently, machine learning (ML) models have started to be used in general practice (GP), although it isn't always clear which algorithms are best or how feature selection (FS) methods affect the results.

While estimating wheat crop production using a number of different FS techniques, Sirsat et al. [31] compared ML and deep learning (DL) algorithms against traditional Bayesian models (in three datasets). They found that compared to the FS method, the prediction algorithm had a bigger effect on model performance. Traditional Bayesian techniques and tree-based ML techniques (random forests and gradient boosting) outperformed all other models in terms of performance. The latter, however, was prone to fitting problems. The only Bayesian FS method used in this work, models built with features selected using Bayes, likewise showed this issue. However, the other three FS techniques generated models with comparable performance but no fitting problems. The authors contend that choosing the prediction algorithm is more important than choosing the FS approach when building highly predictive models as a result. Also, they got to the conclusion that gradient boosting and random forests offer GP models for wheat grain yield that are very reliable and predictive.

**Table 5.** Summary of papers on Genomic and Phenology based wheat yield prediction.

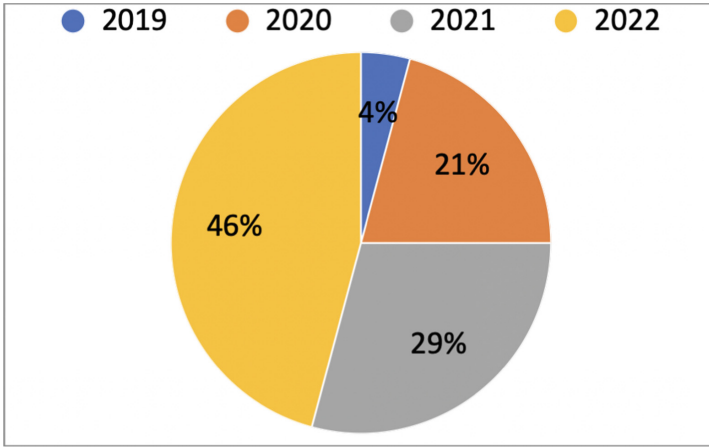
Author	Method	Dataset	Acc	$R^2$	RMSE
Feng et al. [16]	Random Forest	Biophysical data	–	0.62	1000.01 kg/ha
Sirsat et al. [31]	Gradient boosting	Genomic and phenotypic data	–	–	–
Srivastava et al. [32]	CNN	Weather, soil, and crop phenology variables	–	0.65	–

A convolutional neural network model was proposed by Srivastava et al. [32] to anticipate winter wheat yield from environmental and phenological data. The researchers used a dataset of meteorological, soil, and crop phenology characteristics from 1999 to 2019 to investigate the effectiveness of machine learning and deep learning methods for predicting the production of winter wheat. They used RMSE, MAE, and correlation coefficient metrics to assess the prediction power of eight supervised machine learning baseline models. Their findings demonstrated that nonlinear models outperformed linear models in capturing the link between crop yield and input data, including the proposed CNN model, DNN, and XGBoost. For the prediction of winter wheat yield, the suggested CNN model outperformed all previous baseline models (7 to 14% lower RMSE, 3 to 15% lower MAE, and 4 to 50% higher correlation coefficient than the baseline that performed the best across test sets).

## 5 Discussion

We want to discover the many cutting-edge machine learning trends and methodologies used for a precise prediction of wheat output in this literature review. We conducted a systematic evaluation of the literature on research works published within the last four years for this endeavor, and we found 24 articles that addressed the research questions we were looking to answer for Fig. 2. This review’s major research goal is to determine the response to the following research question: **“What cutting-edge machine learning methods have been employed in the last four years to forecast wheat yields?”**

In our review of 24 research articles, we found the Random Forest algorithm and Deep neural network to be the most popular for use in the application of wheat yield prediction with 60% of the reviewed articles utilizing the two machine learning algorithms. These two algorithms are usually applied to remote sensing and time series type of data. The third most used algorithms are families of Gradient boosting algorithms such as AdaBoost, XGboost, and LightGBM,



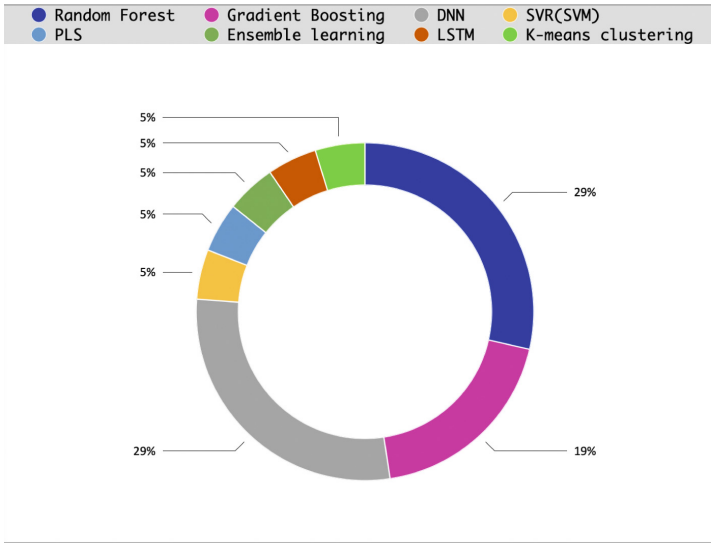
**Fig. 2.** Distribution of reviewed articles with respect to year of publication.

with 19% Fig. 3. We also observed the popularity of the LSTM algorithm for datasets consisting of historical yield prediction records.

As shown in the summary Tables 3, 4, and 5 of the reviewed articles which are presented in Sects. 4.1, 4.2, 4.3; we have identified and discussed the state of the art machine learning techniques used to predict crop yield. This gives the full answer to our primary research question and can give a summarized clue for future researchers in the area.

In addition, our survey explores to address the secondary research questions. Based on this we found that accurate prediction of wheat yields is a crucial factor in minimizing yield loss and ensuring food security. We found these reasons are the key motivation for the increased research interest in the application of machine learning in crop yield prediction. We discovered that the main driver for using ML approaches for wheat yield prediction is to improve prediction accuracy while reducing RMSE losses as we reviewed the majority of studies in the field. To minimize output losses and preserve food security, this is vital for farmers and policymakers to consider when making decisions. Many different dataset types are utilized in various studies. Some of them are satellite data, observational climate data, sub-seasonal to seasonal atmospheric prediction data, genetic data, and phenology data. Based on our survey we found that most studies use only one type of dataset (some papers use only satellite data while others use only climate data or genomic and phenology data). Only a few studies used fused datasets (e.g. satellite image, observational climate, and sub-seasonal to seasonal atmospheric prediction data). Our survey result indicates that merging that dataset would give a better result. In most studies, the primary evaluation metrics used for measuring yield accuracy are  $R^2$  and RMSE values. Based on our survey we found that using a fused dataset and a set of ML models with ensemble learning would give a better result. One of the main challenges identified in this review paper is the lack of a large public dataset, which makes reproducibility and comparison

of model performances very difficult. Most of the data used in the research articles are geographically specific, which decreases the model's generalization performance and complicates models' performance comparisons. In Addition, we observed the authors using a variety of measurement metrics which causes challenges in comparing prediction performances.



**Fig. 3.** Distribution of methodologies used in the reviewed articles.

## 6 Conclusion

For the purpose of this review, we searched through and choose  $N = 24$  journal articles that discuss the various machine-learning approaches used to estimate crop yields over the last five years. This study provides a thorough analysis of yield prediction models that employ machine learning methods. The study issues pertaining to the various machine learning methods/algorithms, the dataset used, the assessment metrics, and the outcomes of each evaluated article are addressed in the paper's presentation. In addition, our survey explores to address the secondary research questions. Based on this we found that accurate prediction of wheat yields is crucial a crucial factor in minimizing yield loss and ensuring food security. We determined that these factors were the primary driving forces for the growing amount of research on the use of machine learning to estimate crop yields. We have observed that dependable and effective options for the task at hand include gradient boosting and Random forest, two machine learning methods. For remote sensing and time series-based wheat production prediction, we also noticed the growing use of deep learning methods like deep convolutional neural networks and LSTMs. In order for the various suggested



machine-learning models to be evaluated and contrasted accurately, the research community needs a single common benchmark dataset.

## References

1. Afework, Y.K., Debelee, T.G.: Detection of bacterial wilt on enset crop using deep learning approach. In: *International Journal of Engineering Research in Africa*. vol. 51, pp. 131–146. Trans Tech Publ (2020)
2. Ali, M., et al.: Coupled online sequential extreme learning machine model with ant colony optimization algorithm for wheat yield prediction. *Sci. Rep.* **12**(1), 1–23 (2022)
3. Bali, N., Singla, A.: Emerging trends in machine learning to predict crop yield and study its influential factors: a survey. *Arch. Comput. Methods Eng.* **29**(1), 95–112 (2022)
4. Biratu, E.S., Schwenker, F., Ayano, Y.M., Debelee, T.G.: A survey of brain tumor segmentation and classification algorithms. *J. Imaging* **7**(9), 179 (2021)
5. Biratu, E.S.S., Schwenker, F., Debelee, T.G.G., Kebede, S.R.R., Negera, W.G.G., Molla, H.T.T.: Enhanced region growing for brain tumor MR image segmentation. *J. Imaging* **7**(2), 22 (2021)
6. Cao, J., Wang, H., Li, J., Tian, Q., Niyogi, D.: Improving the forecasting of winter wheat yields in northern china with machine learning-dynamical hybrid subseasonal-to-seasonal ensemble prediction. *Remote Sens.* **14**(7), 1707 (2022)
7. Chergui, N.: Durum wheat yield forecasting using machine learning. *Artif. Intell. Agric.* **6**, 156–166 (2022)
8. Conradt, T.: Choosing multiple linear regressions for weather-based crop yield prediction with ABSOLUT v1. 2 applied to the districts of Germany. *Int. J. Biometeorol.* **66** 1–14 (2022)
9. Debelee, T.G., Amirian, M., Ibenthal, A., Palm, G., Schwenker, F.: Classification of mammograms using convolutional neural network based feature extraction. In: Mekuria, F., Nigussie, E.E., Dargie, W., Edward, M., Tegegne, T. (eds.) *ICT4DA 2017*. LNICST, vol. 244, pp. 89–98. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-95153-9\\_9](https://doi.org/10.1007/978-3-319-95153-9_9)
10. Debelee, T.G., Kebede, S.R., Schwenker, F., Shewarega, Z.M.: Deep learning in selected cancers’ image analysis-a survey. *J. Imaging* **6**(11), 121 (2020)
11. Debelee, T.G., Schwenker, F., Ibenthal, A., Yohannes, D.: Survey of deep learning in breast cancer image analysis. *Evol. Syst.* **11**(1), 143–163 (2020)
12. Debelee, T.G., Schwenker, F., Rahimeto, S., Yohannes, D.: Evaluation of modified adaptive k-means segmentation algorithm. *Comput. Vis. Media* **5**(4), 347–361 (2019)
13. Elavarasan, D., Vincent, P.M.D.R.: A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. *J. Ambient. Intell. Humaniz. Comput.* **12**(11), 10009–10022 (2021). <https://doi.org/10.1007/s12652-020-02752-y>
14. Fajardo, M., Whelan, B.: Within-farm wheat yield forecasting incorporating off-farm information. *Precision Agric.* **22**(2), 569–585 (2021)
15. Fei, S., et al.: Uav-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. *Precision Agric.* **24**, 1–26 (2022)
16. Feng, P., et al.: Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. *Agric. For. Meteorol.* **285**, 107922 (2020)

17. Filippi, P., et al.: An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agric.* **20**(5), 1015–1029 (2019). <https://doi.org/10.1007/s11119-018-09628-4>
18. Han, J., et al.: Prediction of winter wheat yield based on multi-source data and machine learning in china. *Remote Sens.* **12**(2), 236 (2020)
19. Kamir, E., Waldner, F., Hochman, Z.: Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. *ISPRS J. Photogramm. Remote. Sens.* **160**, 124–135 (2020)
20. Keele, S., et al.: Guidelines for performing systematic literature reviews in software engineering. Technical report, ver. 2.3 EBSE (2007)
21. Lischoid, G., Webber, H., Sommer, M., Nendel, C., Ewert, F.: Machine learning in crop yield modelling: a powerful tool, but no surrogate for science. *Agric. For. Meteorol.* **312**, 108698 (2022)
22. Murakami, K., Shimoda, S., Kominami, Y., Nemoto, M., Inoue, S.: Prediction of municipality-level winter wheat yield based on meteorological data using machine learning in hokkaido, japan. *PLoS One* **16**(10), e0258677 (2021)
23. Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N.H., Islam, N.: A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sens.* **14**(9), 1990 (2022)
24. Nevavuori, P., Narra, N., Lipping, T.: Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* **163**, 104859 (2019)
25. Oikonomidis, A., Catal, C., Kassahun, A.: Deep learning for crop yield prediction: a systematic literature review. *New Zealand J. Crop Hortic. Sci.* 1–26 (2022). <https://doi.org/10.1080/01140671.2022.2032213>
26. Pang, A., Chang, M.W., Chen, Y.: Evaluation of random forests (RF) for regional and local-scale wheat yield prediction in southeast Australia. *Sensors* **22**(3), 717 (2022)
27. Paudel, D., et al.: Machine learning for large-scale crop yield forecasting. *Agric. Syst.* **187**, 103016 (2021)
28. Qiao, M., et al.: Exploiting hierarchical features for crop yield prediction based on 3-d convolutional neural networks and multikernel gaussian process. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **14**, 4476–4489 (2021)
29. Shewry, P.R.: Wheat. *J. Exp. Bot.* **60**(6), 1537–1553 (2009)
30. Shidnal, S., Latte, M.V., Kapoor, A.: Crop yield prediction: two-tiered machine learning model approach. *Int. J. Inf. Technol.* **13**(5), 1983–1991 (2021)
31. Sirsat, M.S., Oblessuc, P.R., Ramiro, R.S.: Genomic prediction of wheat grain yield using machine learning. *Agriculture* **12**(9), 1406 (2022)
32. Srivastava, A.K., et al.: Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Sci. Rep.* **12**(1), 1–14 (2022)
33. Tesfaye, A.A., Awoke, B.G., Sida, T.S., Osgood, D.E.: Enhancing smallholder wheat yield prediction through sensor fusion and phenology with machine learning and deep learning methods. *Agriculture* **12**(9), 1352 (2022)
34. Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., Li, H.: An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the guanzhong plain, pr china. *Agric. For. Meteorol.* **310**, 108629 (2021)
35. van Klompenburg, T., Kassahun, A., Catal, C.: Crop yield prediction using machine learning: a systematic literature review. *Comput. Electron. Agric.* **177**, 105709 (2020). <https://doi.org/10.1016/j.compag.2020.105709>, <https://www.sciencedirect.com/science/article/pii/S0168169920302301>

36. Vanli, Ö., Ahmad, I., Ustundag, B.B.: Area estimation and yield forecasting of wheat in southeastern turkey using a machine learning approach. *J. Indian Soc. Remote Sens.* **48**(12), 1757–1766 (2020)
37. Waldamichael, F.G., Debelee, T.G., Schwenker, F., Ayano, Y.M., Kebede, S.R.: Machine learning in cereal crops disease detection: a review. *Algorithms* **15**(3), 75 (2022)
38. Wang, Y., Zhang, Z., Feng, L., Du, Q., Runge, T.: Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous united states. *Remote Sens.* **12**(8), 1232 (2020)
39. Yang, S., et al.: Integration of crop growth model and random forest for winter wheat yield estimation from UAV hyperspectral imagery. *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.* **14**, 6253–6269 (2021)
40. Zhang, J., Tian, H., Wang, P., Tansey, K., Zhang, S., Li, H.: Improving wheat yield estimates using data augmentation models and remotely sensed biophysical indices within deep neural networks in the Guanzhong plain, PR china. *Comput. Electron. Agric.* **192**, 106616 (2022)
41. Zhong, R., et al.: Detect and attribute the extreme maize yield losses based on spatio-temporal deep learning. *Fundam. Res.* (2022)
42. Zhou, W., Liu, Y., Ata-Ul-Karim, S.T., Ge, Q., Li, X., Xiao, J.: Integrating climate and satellite remote sensing data for predicting county-level wheat yield in china using machine learning methods. *Int. J. Appl. Earth Obs. Geoinf.* **111**, 102861 (2022)