Potato Research (2025) 68:759–792
https://doi.org/10.1007/s11540-024-09753-w

Predicting Potato Crop Yield with Machine Learning and Deep Learning for Sustainable Agriculture

El-Sayed M. El-Kenawy[1] · Amel Ali Alhussan[2] · Nima Khodadadi[3] · Seyedali Mirjalili[4] · Marwa M. Eid[1,5]

Received: 25 April 2024 / Accepted: 7 June 2024 / Published online: 13 July 2024

Abstract

Potatoes are an important crop in the world; they are the main source of food for a large number of people globally and also provide an income for many people. The true forecasting of potato yields is a determining factor for the rational use and maximization of agricultural practices, responsible management of the resources, and wider regions' food security. The latest discoveries in machine learning and deep learning provide new directions to yield prediction models more accurately and sparingly. From the study, we evaluated different types of predictive models, including K-nearest neighbors (KNN), gradient boosting, XGBoost, and multilayer perceptron that use machine learning, as well as graph neural networks (GNNs), gated recurrent units (GRUs), and long short-term memory networks (LSTM), which are popular in deep learning models. These models are evaluated on the basis of some performance measures like mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) to know how much they accurately predict the potato yields. The terminal results show that although gradient boosting and XGBoost algorithms are good at potato yield prediction, GNNs and LSTMs not only have the advantage of high accuracy but also capture the complex spatial and temporal patterns in the data. Gradient boosting resulted in an MSE of 0.03438 and an R^2 of 0.49168, while XGBoost had an MSE of 0.03583 and an R^2 of 0.35106. Out of all deep learning models, GNNs displayed an MSE of 0.02363 and an R^2 of 0.51719, excelling in the overall performance. LSTMs and GRUs were reported to be very promising as well, with LSTMs comprehending an MSE of 0.03177 and GRUs grabbing an MSE of 0.03150. These findings underscore the potential of advanced predictive models to support sustainable agricultural practices and informed decision-making in the context of potato farming.

Keywords Crop yield prediction · Potato · Machine learning · Deep learning · Food security · Sustainable agriculture

Extended author information available on the last page of the article

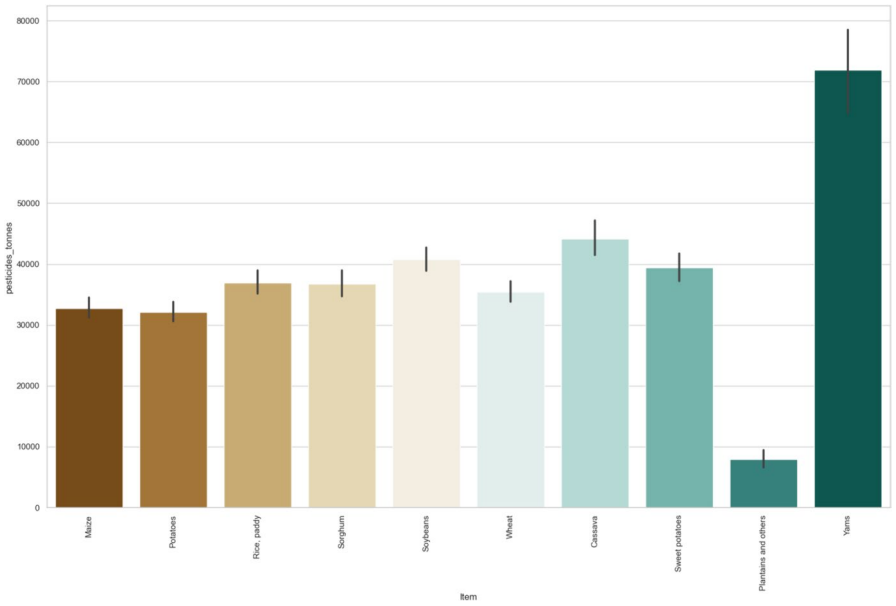

Fig. 1 Comparison between potato production and other agricultural crops

Introduction

Agricultural productivity is foundational to global food security and economic development, underpinning the sustenance and prosperity of societies around the world. As the global population continues to grow at a rapid pace, reaching approximately 9.7 billion by 2050, the challenge of meeting increasing food demands while preserving natural resources is becoming increasingly urgent. Crop yield optimization is essential in this context, as it maximizes the use of limited resources such as water, fertilizers, and land, thereby enabling consistent food production and stabilizing market dynamics (Jayne and Sanchez 2021; Ortiz-Bobea et al. 2021). The conventional approach to crop yield forecasting involved heavily observing historical data and the professional opinions of the agronomists. Such methods usually apply a simple statistical analysis or a linear regression model, which might be fairly accurate in a long-term stable environment but are incapable of accurately including a number of factors that interact with each other in the modern agricultural complex. These mechanisms include the phenomena of modified weather conditions, depletion of soil fertility, the emergence of new pests and diseases, and evolving consumer preferences, which collectively affect the standard predictive yield models (Paudel et al. 2021; Zaki et al. 2023a).

Figure 1 offers a visualization of potato production against other major types of agricultural output. The graph shows the magnitude of potato cultivation in proportion to crops, including wheat, rice, maize, and soybean. Through the delivery of a graphic that indicates the average production volume or yield of potatoes as compared to other crops, you can see that potatoes are among the most significant crops

in the world and contribute greatly to food production. By acknowledging the relationship between potatoes and other crops and disseminating information regarding their relative importance, insights into the new agricultural trends, resource allocation, and food security systems can be drawn.

The importance of potatoes as a globally significant crop can hardly be overstated as they are the major staple food source for many cultures, and they are full of healthy carbohydrates and nutrients. As a semi-aquatic crop, potatoes can be cultivated across different climates and varied types of soils, making a good filling for the niche of the world food supply. But at the same time, careful management is necessary for the potatoes as they are very prone to different diseases like late blight, which can destroy whole harvests. Also, potatoes' susceptibility to water stress and deficiency of some nutrients forces farmers to adhere to precise irrigation and fertilization practices. These technologies (machine learning and deep learning) have the potential to increase the accuracy of potato yield forecast and cultivation strategies. The use of data collected from rainfall records, soil moisture information, and plant health indicators would support this kind of intervention. From this, we get not only increased yields but also more efficient resources and sustainability. However, it all the same increases the productivity and resilience of potato farming (Klompenburg et al. 2020; Cao et al. 2021a, 2021b).

Figure 2 shows the yearly production trend of potatoes, which ranged from 1990 to 2010. The plot serves as a time series depicting how the yield of potatoes has changed over the twenty years, revealing whether the yield is growing or declining and also the periods when the yield is stable. Such actualization needs to be taken into consideration in understanding the dynamics of potato productivity during the historical period, which may have been affected by different factors like strategies in agricultural processes, changes in climatic conditions, pest and disease management, and or variations in utilization and cultivation methods. It is possible to measure the trend, which helps stakeholders evaluate the level of potato production and

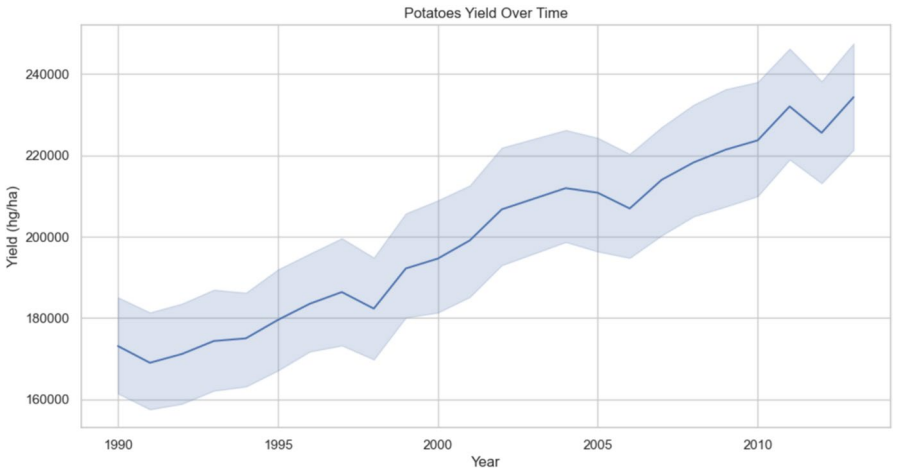

Fig. 2 Potatoes yield over time

determine the periods of rapid changes. Besides this, such historical background could be used for a prognosis, which will allow adjustments in agricultural practices to enhance the yield of potatoes. The figure can also be a means to associate harvest variations with external events, such as technological improvements, policy adjustments, or changes in market demand.

Recently, the emergence of complicated machine learning (ML) (Bali and Singla 2022) and deep learning (DL) (Shook et al. 2021) technologies that can manipulate enormous data sets has brought an evolutionary change to the way crop yields are forecast and give very precise estimates. The methods are based on different types of data collected by satellites and drones, weather and soil sensors, farm management records, etc. ML and DL techniques may use these resources to draw out the abnormalities and relationships buried within the data, which provides more precision and fineness in prediction models. The ability to develop and improve through continuous learning with machine learning models allures among their greatest powers. New data would rather allow models to adjust their forecasts and thus make sure the accuracy is high even for a long time. It helps them to have quick interactive responses to the dynamically changing agricultural environment, not like others that depend on initial data sets that may not alter.

ML and DL models' ability for granularity is another main advantage of them. Such techniques would allow for predictions at the field level or even plant level, which is necessary to conduct interventions of a targeted and precise nature. Farmers may resort to this detailed data and develop their irrigation practices, accuracy of fertilizer application, and control of pests, which, in the long run, will improve their productivity and efficiency (Elavarasan and Vincent 2020; Durai and Shamili 2022). ML/DL models can carry out this synthesis with a vast range of scales and sources of data, giving a unified view of crop states and yield potential. For example, satellites might offer macro pictures of crop patterns and health, and sensors will provide local data for soil moisture and nutrients. ML and DL models would benefit from the integration of these different data sets since the power of the latter is increased (Darwin et al. 2021; Abbas et al. 2020; Shahhosseini et al. 2021).

The import of advanced forecasting models is not confined to individual farms but is dedicated to various components of the agricultural supply chain. Correct yield predictions ensure the right planning for storage, transportation, and sales, which also ensures that the market remains stable and that the risks of food insecurity are mitigated. This generates a positive cycle, which is beneficial to both producers and consumers (Cedric et al. 2022). However, the process of using ML and DL techniques in crop yield prediction is not always free from problems as well. Data sets, which are of high quality and large scale, are essential for training and validation of these models, the sources of which may not be readily available or accessible, as in resource-limited areas. Additionally, the sophistication of these models may also be a drawback because their outputs sometimes cannot be easily understood by users who may need to know the reason for specific predictions given by such models (Wolanin et al. 2020; Kang et al. 2020).

They are keeping in compliance with fairness and bias removal. At the same time, data inputs and model training are essential factors that ensure the reliability and equitable application of ML and DL approaches in agriculture. The prejudices can

be due to uneven data sets. Thus, the predictions may receive a slant that can disadvantage specific groups or regions. Therefore, there is a need to test the validity of models in varying conditions and different data sources (Rashid et al. 2021; Sujatha et al. 2021; Schwalbert et al. 2020). In spite of the difficulties still associated with ML and DL strategies to fuel future yield prediction and agriculture transformation, their potential is enormous. These technologies help farmers adopt better farming methods, conduct agricultural planning beyond the limits of the farm, and shape the policy for sustainable and climate-resilient agricultural systems (Wani et al. 2022; Alibabaei et al. 2021).

A comprehensive comparison of established ML and DL applications for the prediction of crop yield is aimed at this paper to discover what the state-of-the-art approaches can do, reveal their limitations, and predict what could be achieved through further development in this area. The paper presents the results of different models' performance in other countries. It outlines the promising features as well as directions for further development of the models. In addition, this paper will explore the far-reaching consequences of the integration of ML and DL techniques into agriculture, including probable impacts on policy, economics, sustainable development, and ecology. The main purpose is to empower the interested parties with the knowledge they need to use data-based technologies in the future for research, development, and implementation of agricultural technologies (Nevavuori et al. 2020; Anami et al. 2020; Chandraprabha and Dhanaraj 2020).

In sum, this paper aims to contribute to the growing body of knowledge on the use of advanced technologies in agriculture. By identifying effective strategies for improving crop yield prediction and fostering sustainable agricultural practices, the study aspires to facilitate more informed decision-making and the adoption of innovative solutions throughout the farming industry.

The study has the following specific objectives to address gaps in current literature.

1. Comprehensive evaluation of predictive models:
    - Assess various machine learning (KNN, gradient boosting, XGBoost, MLP) and deep learning (GNNs, GRUs, LSTM) models for predicting potato yield, using metrics like MSE, RMSE, and MAE to identify the most accurate models.
2. Handling agricultural data complexity:
    - Utilize advanced ML and DL techniques to manage complex agricultural data, capturing intricate spatial and temporal patterns that traditional linear models cannot handle.
3. Enhancing resource management and food security:
    - Improve agricultural practices and resource management by providing precise yield predictions, supporting sustainable agriculture and food security.
4. Integration of diverse data sources:

- Combine weather conditions, soil properties, historical yield data, and agronomic characteristics to create generalized, robust predictive models applicable across different regions.

5. Model interpretability and usability:
   - Focus on explainable AI to make advanced predictive models accessible and actionable for farmers and policymakers.
6. Contribution to sustainable development
   - Explore the broader impact of ML and DL integration in agriculture, including policy, economic, and ecological implications, to support sustainable development and climate-resilient farming practices.

Related Works

Crop yield prediction is considered critical for agricultural planning, resource management, and food security. Considering that most people eat potatoes and they are crucial to the business, it becomes vital to be able to estimate the amount to be cultivated on each. The ability to accurately forecast yields of potato crops allows those involved, namely farmers, lawmakers, and others, to make wise decisions, utilize resources efficiently, and prevent famine. The works that can be seen in this section are devoted to the prediction of crop returns for potatoes. By considering a range of methods, data sources, and technological innovations, the evolving research environment aimed at improving knowledge and predictive capabilities in the area of potato agriculture is brought to the limelight.

ANNs have been applied effectively in agricultural remote sensing for a number of applications, such as crop classification and crop area estimation. Two types of ANNs are analyzed in the literature (Pandey and Mishra 2017). One is an RBFNN, and the other is a GRNN. They are employed to predict the amount of potatoes that a crop would produce in various forms. Neural networks are trained and evaluated based on leaf area index, biomass, plant height, and other critical indicators of crop performance. Both GRNN and RBFNN prove that they can guess how much potato the crop will produce and correct the guess. The GRNN is a great guesser due to its quick learning and a low spread constant of 0.5. This one is better at this than the RBFNN. The study also provides valuable new perspectives on how to employ ANNs for precise predictions of crop yields in various farming conditions through the comparison of rough surface fields' productiveness with flat ones. To make wise choices in smart farming, you need to be capable of using data captured by sensors placed next to the crops to project the output of such crops in the future. In a previous study (Dubois et al. 2021), the cultivation of potatoes was the focus, but irrigation was essential for the proper growth of crops. The particular problem that was analyzed was soil water potential forecast. In the field of machine learning, supervised learning algorithms are used to accomplish this. Several tests have been conducted over three years on various cases to demonstrate that procedures for feature

selection may be adopted to develop models with appropriate features. Machine learning allows for the correct prediction of the water potential numbers for the future. This gives us valuable information regarding the utilization of machine learning techniques to forecast soil water potential and improve agricultural decisions.

There is a high level of knowledge about how machine learning can estimate yields, and much of it is specialized. However, there are problems when you try to transfer these discoveries to other plants and locations. A full machine learning standard for large-scale crop yield predictions is established (Paudel et al. 2021). It achieves this by combining the agronomic concepts of crop models with machine learning. The baseline is an exemplar of a process that has accuracy, modularity, and utility at the forefront. Among the things used to create the explainable predictors for explainability in crop growth and development are weather, remote sensing, soil records, and the outcomes of field simulations. The process is made up of individual parts, which can be used repeatedly for any crop or country. This means that things can get even better. Through case studies, the yields of five crops grown in three countries were guessed. This demonstrated how competitive the guesses are and what changes could be made to improve them in various conditions. As a first step toward using machine learning in practical applications, setting a standard for large-scale crop growth predictions is a big step in the right direction. In trying to determine the national crop yield, we usually use constructed models of spatial units. This often results in errors and uncertainties. Paudel et al. (Paudel et al. 2022) proposed using regional crop yield predictions in the process of using machine learning to forecast crop yields at different spatial levels, as it will prevent error propagation. A general machine learning process with 35 case studies from nine countries reveals lower NRMSE and error at the regional level than a linear trend model. Usually, when summing up all the regional machine learning projections, their NRMSEs are smaller than those of the country's operating system predictions. Machine learning food yield predictions across regions are precise and demonstrate how areas differ. This is why they are ideal for making big policy decisions at all levels of space.

Crop yield forecasting is a process of predicting a crop yield, taking into consideration a number of factors, including location, weather, soil properties, water tables, and yield from the last year. In a previous study (Shetty et al. 2021), a combination of multilayer perceptron neural network model and random forest regression models are employed to predict four of the major crops grown in the Karnataka region. MAE, MSE, and RMSE are used to train and test the models. The inputs of the system are the weather data and the past yield data from 30 areas in Karnataka. The random forest regression and the multilayer perceptron network are both equally capable of making reasonable guesses. Based on these results, it is obvious that they can be used to predict what happens in the near future. The predictive model is used to estimate crop growth in real time using a simple web application. Mishra et al. (2023) analyzed and forecasted potato production in eight important South Asian nations from 1961 to 2028 using advanced time series and machine learning techniques such as ARIMA, state space, and XGBoost. In validation, ARIMA and state space models are superior to XGBoost. This may imply that it needs to be better tailored. However, every country will have a different pattern of trends. These facts about the future potato supply in South Asia can be used for food security planning and regional agricultural policies.

This study (Abrougui et al. 2019) focuses on the impact of tillage systems on soil characteristics, food production, and the applicability of artificial intelligence models for future forecasts. The most important thing for multiple linear regressions (MLR) and artificial neural networks (ANN) is the organic potato crop yield. According to the study, the tillage practice and soil characterization have a major impact on potato growth. The MLR model can predict crop yield more than the ANN model does. This data allows us to understand the relationship between potato production, types of tillage, and soil properties. Cedric et al. 2022 highlighted concerns about food security on the global level, especially in countries that have population growth challenges, such as agricultural output and the impact of climate change on crops around the world. With the help of machine learning and big data technology, a forecasting model for West African countries is given. The study combines data on climate, weather patterns, yield of crops, and chemicals for six key agricultural commodities. This makes it easier to estimate yields at country level. K-nearest friend, decision trees, and multivariate logistic regression are some models that have proven to work and can help farmers make informed decisions.

Since farming is the basis of a country's economy, Prasad Patnaik and Padhy (Prasad Patnaik and Padhy 2023) examined how machine learning can be used in predicting crop yields and suggesting crops that will do well in particular areas. Several approaches can be applied to analyze both supervised and unsupervised machine learning. The results are checked for accuracy using MSE, MAE, RMSE, and other performance measures after extensive pre-processing of the dataset to ensure accuracy. This comprehensive approach aims to apply machine learning methods intelligently to achieve a strong predictive and productive framework for food yields.

Through our investigation, the interrelated studies shown in this section prove that we are highly reliable at determining potato yields. This reflects how farming and technology are going to share a relationship in the near future. Both sophisticated machine learning methods use remote sensing and the modern versions of traditional approaches that significantly benefit from improvements in meteorological data. All these are directed toward this aim, which is to say the correct word at the given time. The practice of integrating information-based insights and the experience of agronomists becomes increasingly critical as the sector progresses. Data scientists, agronomists, and agricultural communities joined efforts in the literature to help not only make crop yield prediction more accurate but also prevent environmental risks and provide global food security.

Materials and Methods

In this part, the paper explains the data sources, pre-processing techniques, model setups, and evaluation metrics used in the study of crop yield predictions using machine learning and advanced deep learning techniques.

Dataset

The data for this study was sourced from a public dataset on Kaggle, specifically designed for crop yield prediction. This dataset includes seasonal crop yield records along with a diverse array of features that are expected to influence agricultural outcomes. The analysis of crop yield prediction is a significant area in addressing food security and environmental sustainability, which is nowadays the major driver toward the increasing population and climate change. To make decision-making more effective at all levels of agriculture management, from farmers to policymakers who are designing to reduce the impact of climate change strategies, better understanding and accurate forecasting of agricultural yield is essential. For this study, we leveraged the public data from a dataset on Kaggle that provides seasonal crop yield records along with information on a diverse array of factors that are expected to influence agricultural outcomes. This complex data set includes features like weather conditions (e.g., temperature, precipitation, humidity), soil properties, historical yield data for different crops, and perhaps some other agronomic characteristics such as planting and harvesting dates, irrigation practices, and pesticide usage (Crop Yield Prediction Dataset (n.d.)).

The dataset has data covering several years that can be utilized to carry out robust analysis and modeling. The geospatial coverage of the data includes different countries, thus enabling one to have an understanding of various agricultural systems and approaches across different geographical locations. This variety of data supply is fundamental in building predictive models that are much more generalized and flexible in other environments. The dataset has been pre-processed to meet the quality and consistency standards by removing the missing data via the imputation methods if necessary and by discarding the data if there is no possibility of imputation. Models and algorithms routinely require normalization and standardization of features unless the study specifically addresses this issue.

The history yield data is considered to be the target variable for our predictive models, while different other features (for instance, weather, soil, and agronomic practices) are used as inputs. By analyzing the correlations between these traits and the yield of historical seasons, we proposed to create powerful machine-learning instruments for forecasting future crop yields for specific areas. In a nutshell, Kaggle's dataset is a valuable data source for training machine learning models and, consequently, for the study of crop yield prediction across numerous countries within an extended period. Furthermore, such models could reveal trends and evidence that would be used in improving agricultural productivity and sustainability.

Figure 3 below represents the key nations that participated in the research, bearing in mind the diversity of agricultural practices and climatic conditions that they are dealing with. This ensures the multiple perspectives and robustness of our models, which are suitable for different settings.

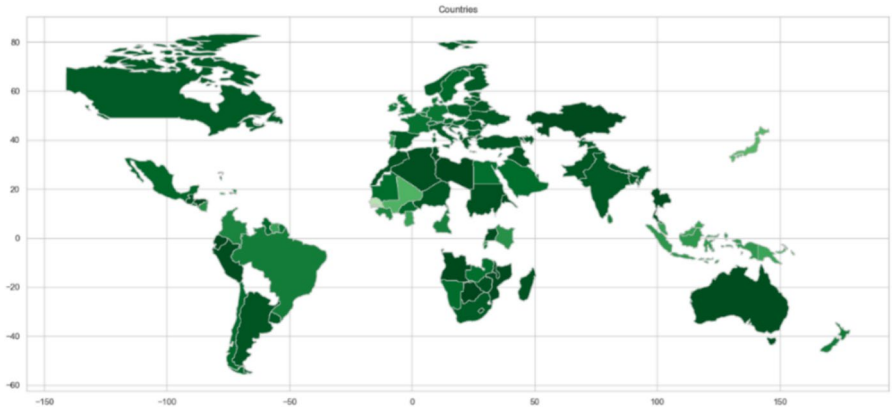

Fig. 3 The countries in which our study was conducted

Data Pre-processing

The stage of data pre-processing is of paramount importance in the process of preparing the data set to facilitate meaningful machine learning and deep learning analyses. Accurate and reliable machine learning algorithms depend on the presentation of clean and transformed data. The data of the present study is a mixture of potato crop data together with data from other crops (Călin et al. 2023). Therefore, data pre-processing is implemented to implement tasks that are responsible for the completion of the study:

- Data cleaning. This step involves identifying and handling missing or incomplete values in the dataset. For instance, missing weather data or yield values may be filled in using imputation techniques, such as mean or median substitution, or removed if they are deemed to introduce bias or uncertainty in the analysis.
- Feature engineering. Features may be added or transformed based on domain knowledge to improve the predictive power of the models. This could include creating new features, such as calculating cumulative precipitation or temperature averages over specific periods, which may be important for crop growth.
- Data standardization and normalization. To ensure compatibility across different features, especially when working with a variety of data types (e.g., weather, soil, and yield data), the dataset must be standardized or normalized. This process ensures that all features contribute equally to model training and prevents the dominance of certain features due to differing scales.
- Outlier detection and handling. Outliers can distort the results of the models and reduce their predictive accuracy. In the data pre-processing phase, outliers are detected using statistical methods such as the z-score or interquartile range (IQR) and handled appropriately, either by transformation or removal, depending on their impact on the data.
- Data splitting. The dataset is split into training, validation, and testing sets to facilitate the development, tuning, and evaluation of predictive models. The

training set is used to build the models, the validation set is used to fine-tune model parameters, and the testing set is used to evaluate the final model's performance.

- Balancing and sampling. In cases where the dataset is imbalanced, such as having significantly more data for one crop over another, balancing techniques such as oversampling or undersampling may be employed to ensure fair and unbiased model training.

In Fig. 4, the potato crop info is shown and contrasted with information about another agricultural crop. This feature is very important for analyzing potato crop data with unique traits and patterns, something that other crops can only offer due to different methods of cultivation, environmental factors, and market demand. Providing these descriptive details to be grasped will guide the processing steps and layer the specific features of potato crop data into the study.

Therefore, data pre-processing tackles these challenging problems, and as a result, the data becomes a strong cornerstone for building accurate and dependable predictive models. The use of cleaned-up data for machine learning and deep learning enables the creation of models that can accurately predict potato crop yield. These forecasting models contribute to decision-making in agriculture.

Data Visualization

Data visualization plays a critical role in understanding the fundamental designs and associations in a data set. Through visualization, we can reveal the inner structure and distribution of data, which enables us to do the modeling and analysis more effectively. This part of the study is about the visualization methods used, which include the potato crop data and its other measurements (Qin et al. 2020).

Figure 5 illustrates the correlation heat map for potato crop measurements. The heatmap displays the pairwise correlations between the different variables in the

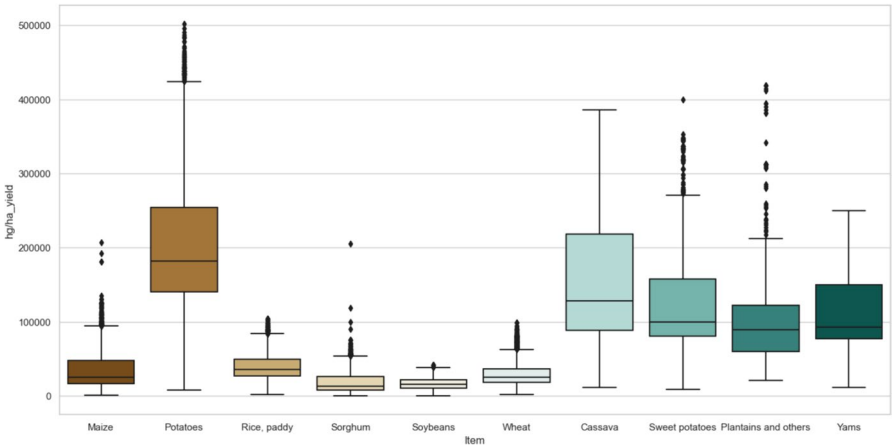

Fig. 4 Outlines found in potato crop data compared to different crops

dataset (such as weather features, including temperature, rainfall, soil properties, and potato yield). According to the heatmap, the color brightness is taken to be the magnitude and direction of the correlation, where warm colors (reds) relate to positive correlation and cool colors (blues) relate to negative correlation. Through this visual approach, the most powerful features associated with potato yield are identified, thereby guiding the feature selection and model generation processes.

Figure 6 shows how potato yield data has been gathered from various countries and divided into regions. According to this visualization, variations in potato yield could reflect variations in agricultural methods, climate differences, or soil variations among regions. Knowledge about the variabilities of these differences can help in creating appropriate approaches to potato farming in a certain area. Furthermore, the graph will be a useful tool in identifying unusual or exceptionally high results that may deserve further examination.

Figure 7 is a hexbin plot showing the correlation between potato production and rainfall. For a hexbin plot, data points are sorted by their proximity to each other into hexagonal bins, which results in a clear view of data density. This visualization may bring to light the existence of such correlations between rainfall and potato yield, i.e., whether high or low rainfall is related to more or less yield. These discoveries can help in the creation of more precise forewarning systems and fine-tuning of irrigation-related activities.

Figure 8 is a histogram analysis of potato yield data, which provides the frequency distribution of potato goods across the entire data set. This graph demonstrates the descriptive statistics of the yield data, specifying the average, dispersion, and skewness. Yield distribution is important because the data mining

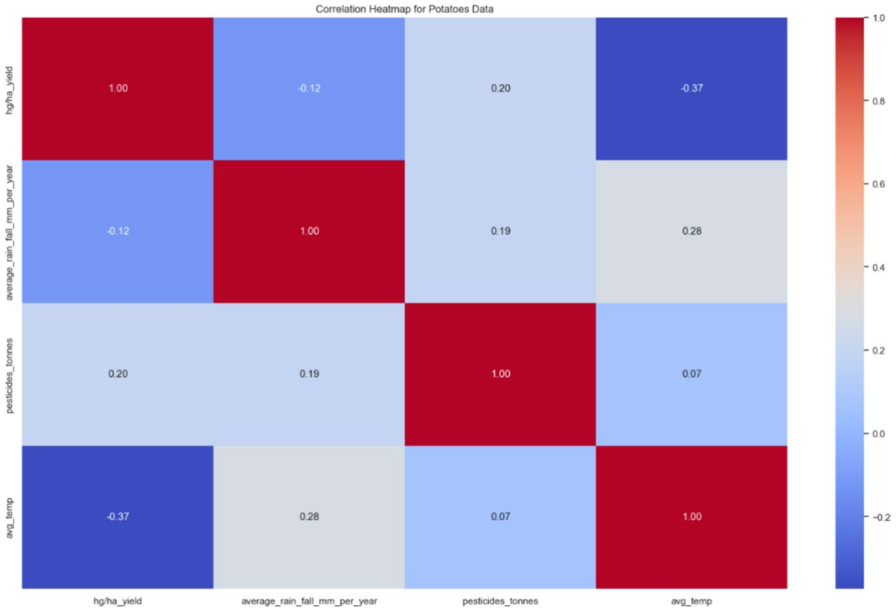

Fig. 5 Correlation heatmap for potatoes data

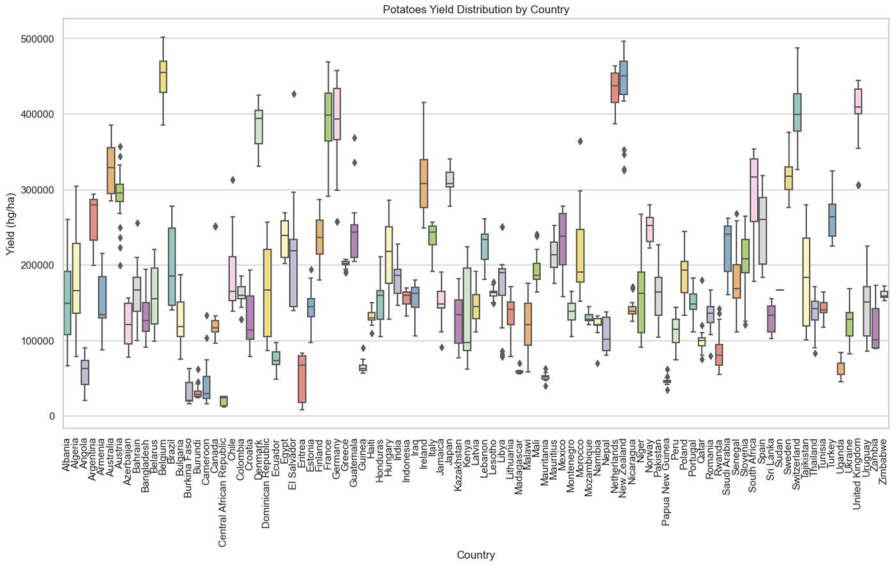

Fig. 6 Potatoes yield distribution by country

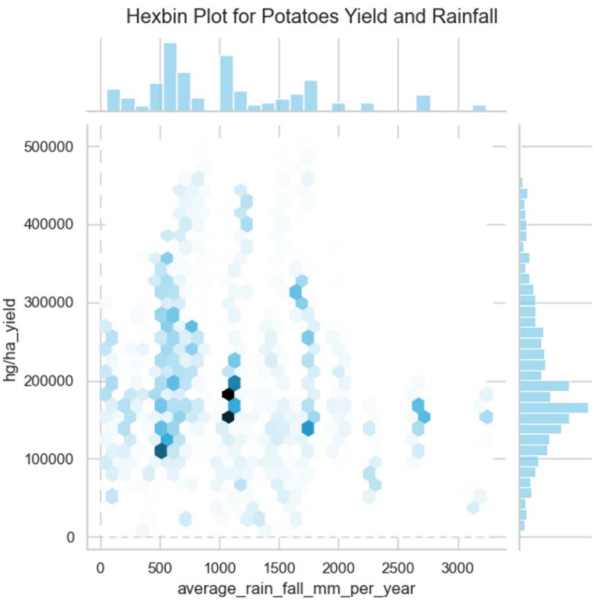

Fig. 7 Hexbin plot for potatoes yield and rainfall

will apply the transformations and choose the model afterward, which will capture more data patterns.

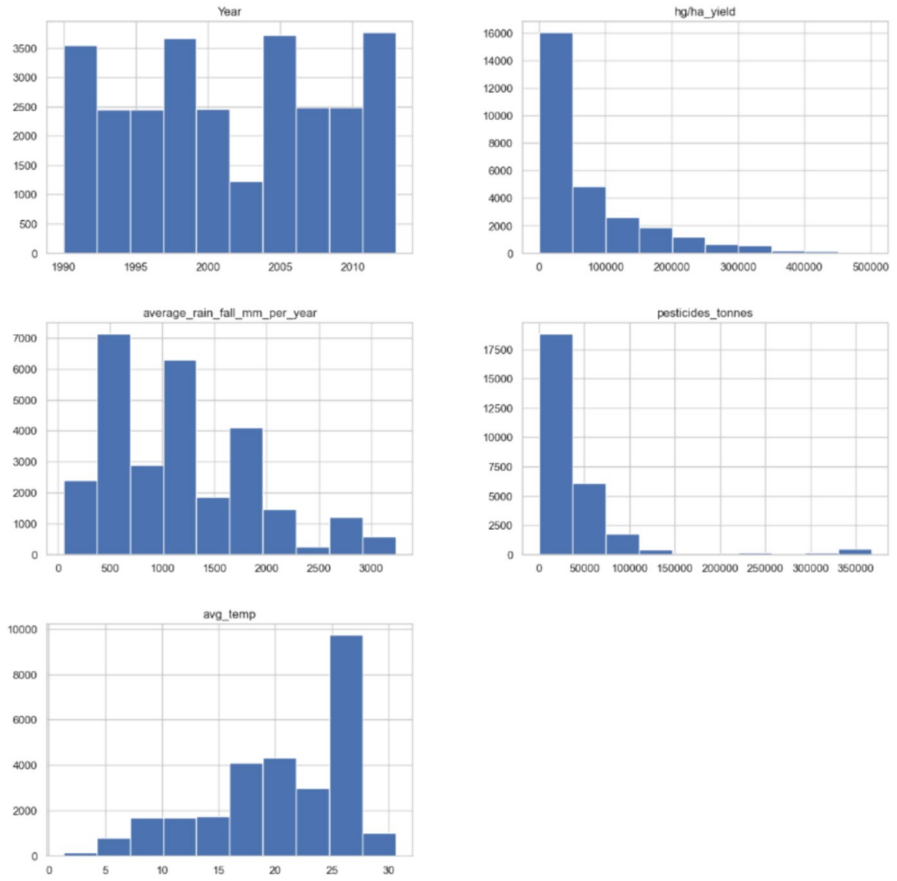

Fig. 8 Histogram analysis for potatoes yield data

These visualizations summarize the set of potato crop data by profitable relationships and trends, which gives an opportunity to run a detailed model. Therefore, this intelligence can be used to create more responsible and precise forecasts about potato production, hence improving the chance of effective decision-making in agricultural management.

Machine Learning Techniques

This section presents the machine learning algorithms utilized in this investigation to forecast crop farming yields. The selected approaches are equipped with a variety of methods, each with its power and capability to handle the different data and prediction tasks. The purpose of these methods is to create predictive models starting from the pre-processed data elements.

K-nearest neighbors (KNN) are quite popular among the simple yet very effective instance-based learning algorithms. In KNN, the prediction for a new data point is determined using a majority vote from its k-nearest neighbors in the training set or a weighted average. The algorithm measures the distance using the most appropriate metric (e.g., Euclidean distance) and then selects the nearest neighbors for the defined data point. KNN algorithm is based on its ability to find clear clustering structures and can predict as soon as it is trained. Nevertheless, it may be sensitive to noise and outliers, and the performance may be affected when the data is multi-dimensional (Zaki et al. 2023b).

Gradient boosting is an ensemble learning technique that develops a model through the successive training of weak learners, which usually are decision trees, and by combining them into a stronger model. The algorithm optimizes the loss function by adding a new model that corrects the errors of the previous ones. The process of this iteration continues until a stopping criterion is reached, such as the number of iterations or a specific level of performance. Gradient boosting is effective in predicting and can process different kinds of input data. Nevertheless, that calls for the delicate tuning of the hyperparameters to prevent overfitting (Abdelmalak et al. 2023).

XGBoost (extreme gradient boosting) is one of the most effective imports in gradient boosting, and it has lots of tweaks for output presentation and efficiency. It ends up using regularization in order to fight overfitting and also makes sure parallelization occurs in order to optimize training time. One of its characteristics, XGBoost, is that it has a collection of objectives, functions, and performance measures that are designed for various tasks. This algorithm is exactly what machine learning competitions need, as it provides and delivers accuracy, speed, and versatility all at once (Noorunnahar et al. 2023).

Multilayer perceptron (MLP) is a category of neural networks that have a layered neural structure in which there are a number of layers of neurons. Neurons in the layer receive signals from previous layers and apply a non-linear activation function, which results in producing output. This output is used as input to the succeeding layers. The lower layer of the model is designed to make the output of predictions from the model. MLPs can recognize and explain extremely intricate, non-linear maneuvers in their possessed data through the optimization of SGD and BP algorithms that adjust their performance. MLPs are known for their high flexibility and performance, especially when working with large datasets and complex patterns. Still, they do have a tricky part: maintaining an equilibrium on different parameters (Ahmed 2023).

These machine learning methods consist of a plethora of methods that exhibit the most benefits for particular crop types and general preferences of the task. In addition to comparing and evaluating model(s), we are working on getting the best present algorithm(s) working for the study.

Deep Learning Techniques

Deep learning algorithms utilize neural network structures that consist of multiple layers to represent naturally the interrelations in data and to achieve highly efficient predictions:

Graph neural networks (GNNs) constitute an engineering family of neural networks that perform well when data are represented as graphs. Crop yield prediction by GNNs constitutes a dependency structure that utilizes the modeling of relationships, whereas geographic localities and crop types represent data points. The graph's composition is made up of nodes representing the bear elements (such as a farm or a crop). At the same time, the boundaries mark the relationships between them (such as proximity, trade, or resemblance). GNN computes the features of the nodes as an edge among the graphs, which provide local information and further amplify across the layer to capture implicit complex relationships. The model can be made more sophisticated to deal with the field of agricultural relations, which is bursting with intricacy (Fan et al. 2022).

Gated recurrent units (GRUs), which are a kind of recurrent neural network (RNN), are discussed as a solution to the problem of the inability to learn from sequential data like weather patterns and crop growth over time. GRUs possess gate mechanisms (update and reset gates) for signal flow control that allow the model to decide what it should keep from the past and throw away presumably irrelevant. The recurrent connection is for this RNN's long-term dependence and time series remembrance. GRN is especially great in serving the purposes of future crop yield prediction because it is always necessary to obtain historical weather patterns and sequential data to decide to prepare for the coming crop season (Jin et al. 2020).

Long short-term memory networks (LSTMs) constitute yet another type of RNN designed to handle sequential information. On top of that, LSTMs and GRUs utilize gating mechanisms (input, output, and forget gates) and govern interactions throughout the entire network. In this manner, LSTMs can recall exactly the comprehensive memory of meaningful information and associate it with longer sequences of the data and long-term dependencies. LSTMs are well suited for time series analysis, as this type of data contains information that has a movement in time from the past to the future. These models also carry information on the previous and future time frames. These give great insight into historical trends (Mateo-Sanchis et al. 2023).

These deep learning techniques that are useful with complication and high-dimensional or sequential have the potential to become perfect prediction tools for crop yields. This research is based on the use of GNNs, GRUs, and LSTMS, which are capable of strong performance in modeling and showing their true, precise, and explainable predictions. The models that were developed will be used in the later decision-making process and will be part of the agricultural practices.

Hyperparameter Tuning Process

Hyperparameter tuning is a crucial step in developing effective machine learning and deep learning models. The process involves selecting the optimal set of hyperparameters that improve model performance. Here, we discuss the hyperparameter tuning process for the specific models used in this study:

1. **K-nearest neighbors (KNN)**
   - Hyperparameters
     - Number of Neighbors (k)
     - Distance Metric (e.g., Euclidean, Manhattan)
   - Tuning Process:
     - Grid Search: A range of k values (e.g., 1 to 20) was evaluated using grid search. Different distance metrics were also tested.
     - Cross-Validation: 5-fold cross-validation was used to assess the performance of each combination.
   - Impact:
     - The optimal k and distance metric were chosen based on minimizing error metrics (e.g., MSE). Smaller k values tended to capture more local patterns, while larger k values smoothed predictions.
2. **Gradient Boosting**
   - Hyper-parameters:
     - Number of Trees
     - Learning Rate
     - Maximum Depth of Trees
   - Tuning Process:
     - Grid Search and Random Search: Both grid search and random search were employed to explore combinations of the number of trees, learning rate (e.g., 0.01, 0.1, 0.2), and tree depth (e.g., 3 to 10).
     - Cross-Validation: 5-fold cross-validation was used to evaluate the performance.
   - Impact:

- Optimal parameters were selected to balance the bias-variance trade-off. A lower learning rate with a higher number of trees typically improved accuracy but increased computational time.

3. **XGBoost (Extreme Gradient Boosting)**

- Hyper-parameters:

  - Number of Trees
  - Learning Rate
  - Maximum Depth of Trees
  - Subsample Ratio
  - Colsample_bytree

- Tuning Process:

  - Bayesian Optimization: Bayesian optimization was used to efficiently navigate the hyper-parameter space.
  - Grid Search: Follow-up grid search was conducted for fine-tuning the bestperforming ranges.
  - Cross-Validation: 5-fold cross-validation was employed.

- Impact:

  - Regularization parameters (e.g., alpha, lambda) helped prevent overfitting. The combination of subsample and colsample_bytree parameters improved generalization by controlling the randomness in model training.

4. **Multilayer Perceptron (MLP)**

- Hyper-parameters:

  - Number of Hidden Layers
  - Number of Neurons per Layer
  - Activation Function (e.g., ReLU, Sigmoid)
  - Learning Rate
  - Batch Size

- Tuning Process:

  - Grid Search: A grid search over the number of layers (e.g., 1 to 5), neurons per layer (e.g., 10 to 100), and activation functions.
  - Random Search: Random search for learning rates (e.g., 0.001 to 0.1) and batch sizes (e.g., 16, 32, 64).
  - Cross-Validation: 5-fold cross-validation was used to select the best configuration.

- Impact:

- The choice of activation function and the number of neurons directly affected the model's ability to capture non-linear relationships. A balanced network depth and size helped manage overfitting and training time.

5. **Graph Neural Networks (GNNs)**

- Hyper-parameters:
  - Number of Layers
  - Learning Rate
  - Batch Size
  - Number of Neurons per Layer
- Tuning Process:
  - Grid Search: Grid search for the number of layers (e.g., 2 to 6) and neurons per layer.
  - Random Search: Random search for learning rates and batch sizes.
  - Cross-Validation: Used for performance evaluation.
- Impact:
  - Deeper networks with more layers were able to capture complex spatial relationships but required careful regularization to avoid overfitting.

6. **Gated Recurrent Units (GRUs)**

- Hyper-parameters:
  - Number of Layers
  - Number of Units per Layer
  - Learning Rate
  - Batch Size
- Tuning Process:
  - Grid Search: Explored number of layers (e.g., 1 to 3) and units per layer (e.g., 50 to 200).
  - Random Search: Applied for learning rates and batch sizes.
  - Cross-Validation: 5-fold cross-validation was used.
- Impact:
  - More layers and units enhanced the ability to capture temporal dependencies but required regularization techniques to prevent overfitting.

7. **Long Short-Term Memory Networks (LSTMs)**

- Hyper-parameters:

- Number of Layers
- Number of Units per Layer
- Learning Rate
- Batch Size

- Tuning Process:

  - Grid Search: Grid search for the number of layers and units per layer.
  - Random Search: Random search for learning rates and batch sizes.
  - Cross-Validation: Employed to determine the best configurations.

- Impact:

  - LSTMs benefited from a higher number of units per layer for capturing longterm dependencies but required careful tuning of learning rates to ensure stable training.

The hyperparameter tuning process involved a combination of grid search, random search, and Bayesian optimization to find the optimal settings for each model. Cross-validation ensured that the selected hyperparameters generalized well to unseen data. This thorough tuning process significantly improved model performance, robustness, and reliability, thereby strengthening the study's validity and practical applicability in agricultural yield prediction.

Experimental Results

This section reveals the results of the feasibility analysis of various machine learning and deep learning approaches applied for crop yield forecasting. The evaluation criteria: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean bias error (MBE), Pearson's correlation coefficient (R), coefficient of determination (R^2), relative root mean squared error (RRMSE), Nash–Sutcliffe efficiency (NSE), Willmott index (WI). These metrics enable the measurement, comparison, and evaluation of various models in terms of accuracy, reliability, and efficiency of the models.

Justification for the Selection of Specific Machine Learning Models

1. **K-nearest neighbors (KNN):**

   - **Justification**: KNN is chosen for its simplicity and effectiveness in handling small to medium-sized datasets. It works well for identifying patterns and trends in the data by using a distance metric (e.g., Euclidean distance) to find the nearest neighbors. KNN does not require an intensive training process and can adapt quickly to new data.

- **Advantages**: Simple, easy to implement, and effective for exploratory analysis.
- **Use case**: Suitable for understanding the structure and distribution of the data and for initial benchmark comparisons.

2. **Gradient boosting:**

- **Justification**: Gradient boosting builds strong predictive models by combining the strengths of multiple weak learners. It incrementally improves the model by focusing on correcting the errors of the previous learners. This iterative approach allows for high accuracy and robust performance.
- **Advantages**: High predictive accuracy, effective for various input data types, and good at capturing complex relationships.
- **Use case**: Ideal for tasks requiring detailed and accurate predictions, especially in scenarios with complex data interactions.

3. **XGBoost (extreme gradient boosting):**

- **Justification**: XGBoost enhances the basic gradient boosting technique by adding regularization to prevent overfitting and parallelization for faster training. It is known for its scalability, efficiency, and high performance in predictive tasks.
- **Advantages**: Regularization to combat overfitting, fast training through parallelization, and high accuracy.
- **Use case**: Suitable for large-scale data analysis and situations where computational efficiency and high precision are required.

4. **Multilayer perceptron (MLP)**:

- **Justification**: MLPs are neural networks that can model complex, non-linear relationships in the data. They consist of multiple layers of neurons that transform the input data through non-linear activation functions, making them highly flexible and capable of handling diverse data patterns.
- **Advantages**: High flexibility, ability to model non-linear interactions and strong performance on complex datasets.
- **Use case**: Effective for modeling intricate patterns and dependencies, especially in datasets with non-linear relationships.

Justification for the selection of specific deep learning models:

1. **Graph neural networks (GNNs):**

- **Justification**: GNNs are designed to work with data structured as graphs, making them ideal for capturing spatial dependencies and relationships between different data points. They leverage the graph structure to aggregate information from neighboring nodes, which is particularly useful for geographical and relational data.
- **Advantages**: Excellent for spatial data, ability to capture complex relationships and strong performance on tasks involving networked data.

- **Use case**: Suitable for analyzing spatial data and complex interactions in agricultural contexts, such as relationships between different farms or crop types.

2. **Gated recurrent units (GRUs):**

- **Justification**: GRUs are a type of recurrent neural network (RNN) that effectively handle sequential data. They include gating mechanisms to control the flow of information and retain relevant historical data, making them suitable for time series analysis and predictions based on historical trends.
- **Advantages**: Efficient handling of sequential data, reduced complexity compared to other RNNs, and effective in capturing temporal dependencies.
- **Use case**: Ideal for tasks involving time series data, such as predicting future crop yields based on weather patterns and historical growth cycles.

3. **Long short-term memory networks (LSTMs):**

- **Justification**: LSTMs are an advanced type of RNN designed to capture long-term dependencies in sequential data. They use input, output, and forget gates to manage the flow of information, preventing the vanishing gradient problem and allowing for the retention of long-term information.
- **Advantages**: Excellent at handling long-term dependencies, effective for time series analysis, and robust against vanishing gradient issues.
- **Use case**: Best suited for modeling long-term trends and making predictions based on extended historical data, such as multi-year crop yield patterns.

The selection of these specific machine learning and deep learning models is based on their strengths and suitability for handling the diverse and complex nature

Table 1 Criteria for evaluating regression result

Metric	Formula
RMSE	$\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\widehat{V}_n - V_n\right)^2}$
RRMSE	$\frac{RMSE}{\sum_{n=1}^{N}\widehat{V}_n} \times 100$
MAE	$\frac{1}{N}\sum_{n=1}^{N}\left\lvert\widehat{V}_n - V_n\right\rvert$
MBE	$\frac{1}{N}\sum_{n=1}^{N}\left(\widehat{V}_n - V_n\right)$
NSE	$1 - \frac{\sum_{n=1}^{N}\left(V_n - \widehat{V}_n\right)^2}{\sum_{n=1}^{N}\left(V_n - \widehat{V}_n\right)^2}$
WI	$1 - \frac{\sum_{n=1}^{N}\left\lvert\widehat{V}_n - V_n\right\rvert}{\sum_{n=1}^{N}\left\lvert V_n - \overline{V}_n\right\rvert + \left\lvert\widehat{V}_n - \overline{V}_n\right\rvert}$
R^2	$1 - \frac{\sum_{n=1}^{N}\left(V_n - \widehat{V}_n\right)^2}{\sum_{n=1}^{N}\left(\sum_{n=1}^{N}V_n\right) - V_n\right)^2}$
r	$\frac{\sum_{n=1}^{N}\left(\widehat{V}_n - \widehat{V}_n\right)\left(V_n - \overline{V}_n\right)}{\sqrt{\left(\sum_{n=1}^{N}\left(\widehat{V}_n - \overline{\widehat{V}}_n\right)^2\right)\left(\sum_{n=1}^{N}\left(V_n - \overline{V}_n\right)^2\right)}}$

of agricultural data. By leveraging the unique advantages of each model, the study aims to achieve high accuracy and robustness in potato yield prediction, ultimately supporting more informed and sustainable agricultural practices.

Performance Metrics

Table 1 shows the criteria used for evaluating the regression results, encompassing various metrics to comprehensively assess model performance. The root mean squared error (RMSE) measures the square root of the average squared differences between predicted and actual values, providing insight into the magnitude of prediction errors. The relative root mean squared error (RRMSE) normalizes RMSE by the sum of actual values, expressed as a percentage, allowing for comparisons across different scales. Mean absolute error (MAE) calculates the average of the absolute differences, offering a straightforward measure of accuracy less sensitive to outliers. Mean bias error (MBE) measures the average bias, indicating tendencies to overestimate or underestimate. Nash–Sutcliffe efficiency (NSE) compares the predictive skill to the mean of observed data, with values closer to 1 indicating better performance. Willmott's index of agreement (WI) evaluates the degree of agreement between predicted and observed values. The coefficient of determination (R^2) indicates the proportion of variance predictable from the independent variables. Lastly, Pearson's correlation coefficient (r) measures the linear correlation between predicted and actual values, with values closer to 1 or -1 indicating stronger relationships. These metrics ensure a thorough evaluation of model accuracy, reliability, and predictive skill.

Machine Learning Techniques Results

Table 2 shows the revealed efficiency indices of several machine learning methods that have been used to predict crop yield in the study. The metrics include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean bias error (MBE), Pearson's correlation coefficient (R), coefficient of determination (R^2), relative root mean squared error (RRMSE), Nash–Sutcliffe efficiency (NSE), Willmott's index (WI), and fitted time. The metrics illustrate how precise and unbiased the models are in their predictions and also reveal the efficiency of predictions concerning a particular model.

In Fig. 9, the MSE charts show the disparity among the assorted machine learning techniques. The MSE calculation results in the sum of the quadratic differences between predicted and actual values, which is used as the accuracy indication. The opposite is true; the lower values indicate better predictive performance.

Figure 10 illustrates a violin plot that depicts the distribution of prediction errors for each machine-learning model in a graphical way. The plot allows for the assessment of error variability between models as well as the positioning of helpful models within this variety, which helps to understand their degree of accuracy and reliability.

Table 2 Machine learning techniques results

Models	MSE	RMSE	MAE	MBE	R	R^2	RRMSE	NSE	WI	Fitted time
K-nearest neighbors (KNN)	0.03437	0.18540	0.13977	0.02242	0.55489	0.30791	40.39800	0.24715	0.61173	0.14118
Gradient boosting	0.03438	0.18542	0.15122	0.06800	0.70120	0.49168	40.40220	0.24699	0.57994	0.37813
XGBoost	0.03583	0.18928	0.14835	0.06810	0.59250	0.35106	41.24392	0.21529	0.58789	7.40334
Multilayer perceptron	0.04125	0.20309	0.16177	0.07496	0.61145	0.37388	44.25308	0.09660	0.55061	9.56434

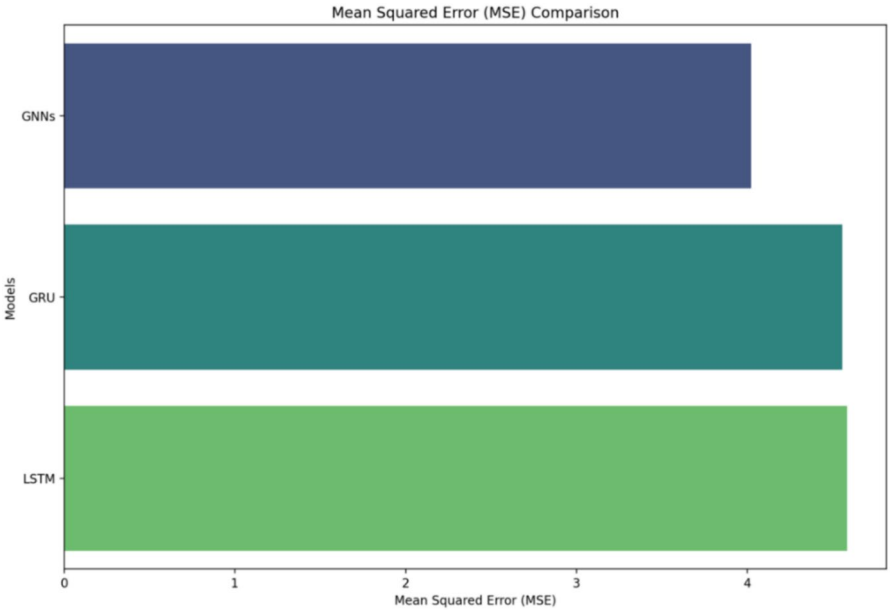

Fig. 9 Mean squared error (MSE) comparison for machine learning techniques

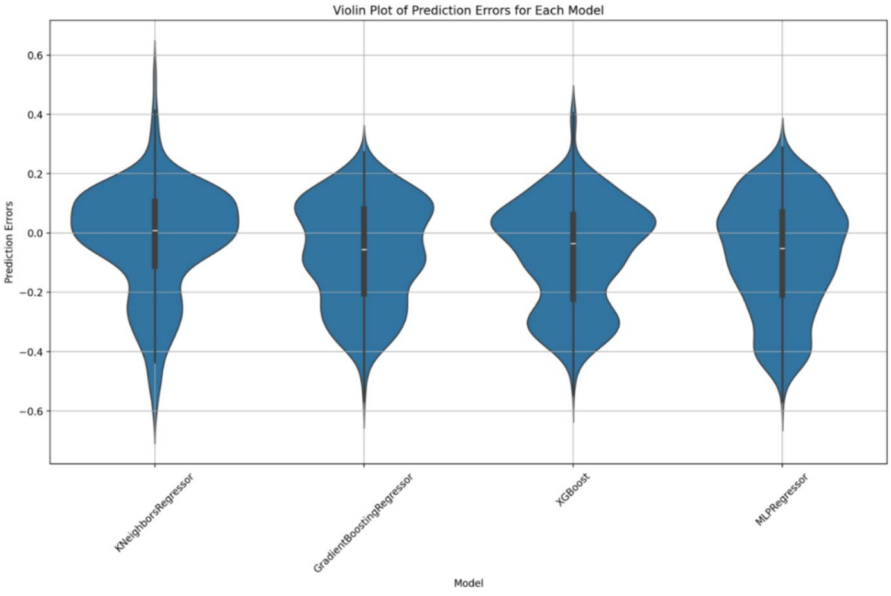

Fig. 10 Violin plot of prediction errors for machine learning model

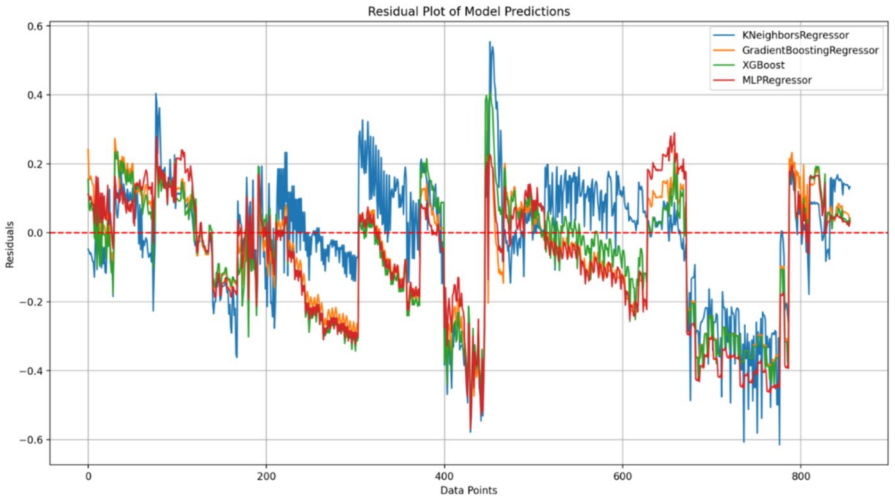

Fig. 11 Residual plot of machine learning model predictions

Figure 11 shows the residual plot of machine learning model predictions. Residuals convey the difference between anticipated and in-place outcomes, and a properly fitted model would indicate a random diffusion around the zero line. The plot helps us to investigate the model's functioning and detect patterns or biases in its predictions.

The results of these experiments have given us some knowledge about how machine learning methods are better at predicting crop yields. These insights can also optimize the models for real-world situations.

Deep Learning Techniques Results

In this section, we will present the findings of our deep-learning approach to crop yield prediction. Table 3 below gives the GNNs that are compared to GRUs and LSTMs. Metrics comprise mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MIE), mean bias error (MBE), Pearson correlation coefficient (R), coefficient of determination (R^2), relative root mean squared error (RRMSE), Nash–Sutcliffe efficiency (NSE), Willmott's index (WI), and fitted.

Figure 12 summarizes MSE differences among all deep learning techniques in this research, including GNNs, GRUs, and LSTMs as model comparisons. MSE measures the square of the mean squared error between the predicted and the actual values, which renders a measure of the precision. Smaller values in the MSE coefficient mean that the model is more precise and accurate in terms of prediction.

By means of this visual comparison, the efficiency of different deep learning approaches can be assessed and ranked, thus providing research and practice communities with a tool to identify the most productive approach to crop yield prediction. Such a comparison, in fact, cannot be overstated in guiding the correct selection

Table 3 Deep learning techniques results

Models	MSE	RMSE	MAE	MBE	R	R^2	RRMSE	NSE	WI	Fitted time
Graph neural networks (GNNs)	0.02363	0.15371	0.12004	0.03614	0.71916	0.51719	33.49355	0.48250	0.66655	0.00168
Gated recurrent units (GRUs)	0.03150	0.17747	0.13990	0.05466	0.62228	0.38723	38.67003	0.31017	0.61137	0.03136
Long short-term memory networks (LSTMs)	0.03177	0.17825	0.14107	0.05146	0.66886	0.44737	38.84123	0.30405	0.60813	0.05049

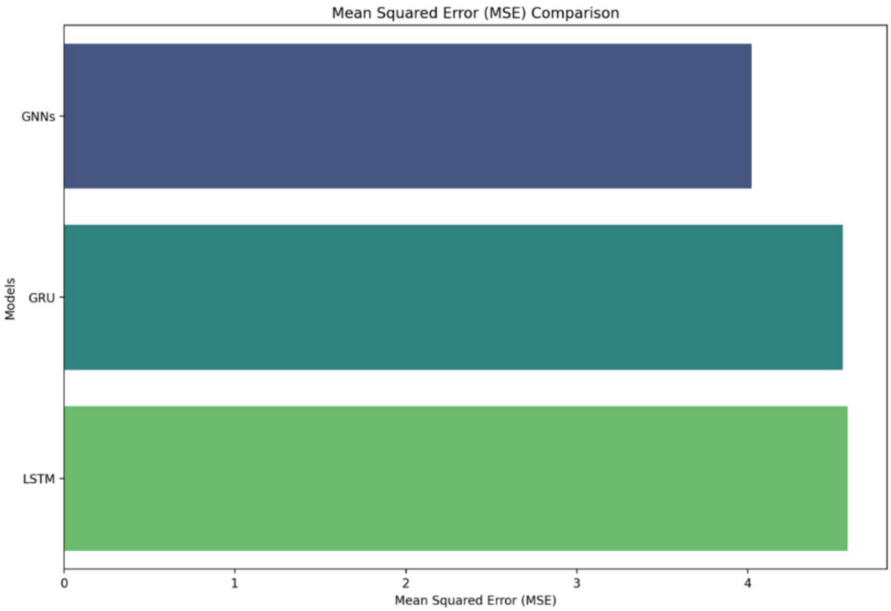

Fig. 12 Mean squared error (MSE) comparison for deep learning techniques

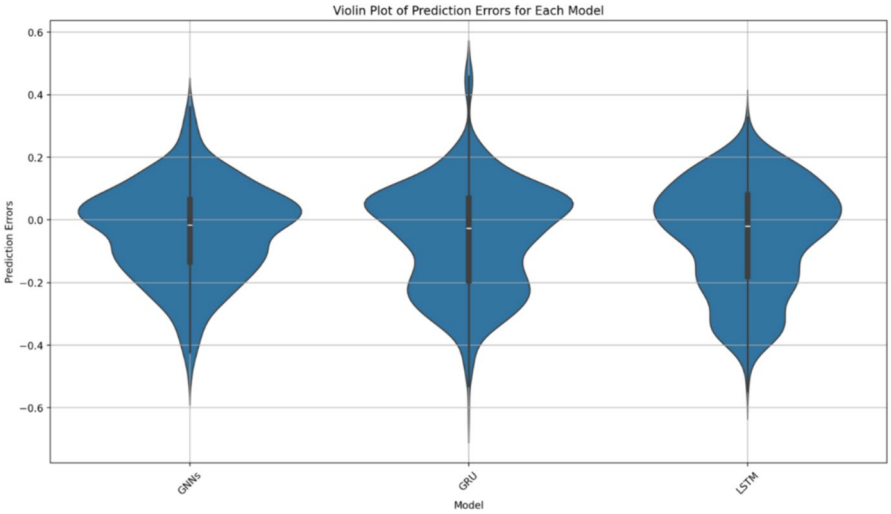

Fig. 13 Violin plot of prediction errors for deep learning model

and application of deep learning models in reality because it may be crucial for supply and demand projections, and this may lead to appropriate planning and resource allocation.

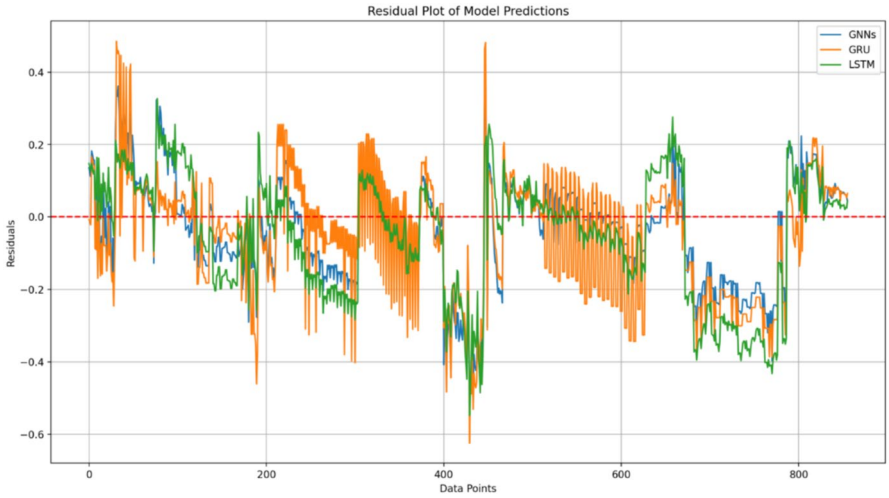

Fig. 14 Residual plot of deep learning model predictions

Figure 13 illustrates a concise visualization of deep learning model prediction errors for each model through which violin plots are used. The violin plot carries the merger of the properties from both the boxplot and the kernel density plot onward so that the distribution and dispersion of data points can be represented in a highly detailed manner for each model. Each one of the violins depicts the density of prediction errors for different levels of errors on the *y*-axis, and at the same time, the shape and the symmetry of each violin gives the nature of the error.

The violin plots are useful tools as you can easily see how widely the predictive errors are spread from each model. This chart can be examined to determine the confidence and stability of the prediction results. Furthermore, it can detect any bias or outliers in the data, which will signal a need for further model fine-tuning and optimization phases.

Figure 14 depicts the residual plot for predictions that the deep learning models made. Residual values are the differences between the observed values and the model's prediction, and the residual plot examines the dependence of the model's predictions on the input data. Ideally, the plot of actual and predicted events should show residuals scattered across the zero line, such that the models do not prejudice the predictions and do not systematically over- or under-predict.

The plot provides a platform to visually check whether the residuals cluster or systematically deviate from the central tendency of the model. Such deviation may convey the possibility of violations in the underlying assumptions or a need for more feature engineering. Uncovering the hidden patterns is obligatory for achieving model performance and making them more precise. In conclusion, the experiment plot confirms model strengths and weaknesses in crop yield prediction under differential climatic scenarios.

Conclusion and Future Direction

In the study, we sought to find the best machine learning and deep learning tools for forecasting crop yields. The reporting data confirm the promise of those methods to enhance prediction accuracy and to give additional information for planning and making the right decisions. Both ML and DL models provide distinct strengths, having a model based on different fatigue detection mechanisms used besides the issue of model precision and time needed to compute. With the help of empirical data, we determined the adequate models for better forecasting output from crops by taking into account the ecological aspects. It will help to develop techniques for increased agricultural productivity.

One of the techniques based on machine learning algorithms used for crop yield forecasting was K-nearest neighbors; another one was gradient boosting; some stopped at XGBoost and multilayer perceptron to verify the accuracy and reliability of the information that they obtained. The gradient boosting method and XGBoost, in particular, were observed to have a good performance, with the model performing well across different evaluation metrics and presenting a high predictive accuracy and robustness.

Deep neural networks, such as graph neural networks, gated recurrent units, and long short-term memory networks, were also considered. These models can recognize intricate details in the data and improve the precision of predictions. GNNs were perceived to be productive for spatial aspects of data as they could capture the complex relationships and interconnections among several data points. In contrast, GRUs and LSTMs distinguished themselves as they displayed strong performance on temporal dependencies and patterns of sequential data.

While the results of this study are promising, there are several areas for future research and development:

- Model optimization and fine-tuning: Further optimization of hyperparameters and model architectures could lead to even greater improvements in predictive performance. Techniques such as automated machine learning (AutoML) could be employed to streamline this process.
- Integration of additional data sources: Incorporating more diverse data sources such as satellite imagery, remote sensing data, or market trends could enhance the models' capabilities and lead to more accurate predictions.
- Explainable AI: As these advanced models become increasingly complex, developing methods to improve the interpretability and explainability of the predictions is essential for practical adoption by farmers and policymakers.
- Real-time predictions: Exploring the feasibility of real-time or near-real-time predictions can provide valuable insights for dynamic decision-making in agriculture.
- Scalability and generalization: Research into the scalability of these models to larger datasets and different geographic regions will help ensure their broader applicability and impact.

- Collaborative research: Continued collaboration between data scientists, agricultural experts, and policymakers is critical to developing effective, practical, and sustainable solutions for crop yield prediction.

Finally, the utilization of machine learning and deep learning technologies for crop yield prediction would play a vital role in the advancement of modern agriculture and the resolution of hunger problems worldwide. More research in these areas and a desire for improvement and optimization will contribute to the development of precision farming and the efficient management of agricultural resources.

Acknowledgements Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R 308), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Data Availability Data are in a repository as public data at https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing interests The authors declare no competing interests.

References

Abbas F, Afzaal H, Farooque AA, Tang S (2020) Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy 10(7):7. https://doi.org/10.3390/agronomy10071046

Abdelmalak MES, Gaber KS, Ahmed MA, OubeBlika N, Zaki AM, Eid MM (2023) BER-XGBoost: pothole detection based on feature extraction and optimized XGBoost using BER Metaheuristic Algorithm. J Artif Intell Metaheuristics 6(2):46–55. https://doi.org/10.54216/JAIM.060205

Abrougui K, Gabsi K, Mercatoris B, Khemis C, Amami R, Chehaibi S (2019) Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). Soil Till Res 190:202–208. https://doi.org/10.1016/j.still.2019.01.011

Ahmed S (2023) A software framework for predicting the maize yield using modified multi-layer perceptron. Sustainability 15(4):4. https://doi.org/10.3390/su15043017

Alibabaei K, Gaspar PD, Lima TM (2021) Crop yield estimation using deep learning based on climate big data and irrigation scheduling. Energies 14(11):11. https://doi.org/10.3390/en14113004

Anami BS, Malvade NN, Palaiah S (2020) Deep learning approach for recognition and classification of yield affecting paddy crop stresses using field images. Artif Intell Agric 4:12–20. https://doi.org/10.1016/j.aiia.2020.03.001

Bali N, Singla A (2022) Emerging trends in machine learning to predict crop yield and study its influential factors: a survey. Arch Comput Methods Eng 29(1):95–112. https://doi.org/10.1007/s11831-021-09569-8

Călin AD, Coroiu AM, Mureşan HB (2023) Analysis of preprocessing techniques for missing data in the prediction of sunflower yield in response to the effects of climate change. Appl Sci 13(13):13. https://doi.org/10.3390/app13137415

Cao J, Zhang Z, Luo Y, Zhang L, Zhang J, Li Z, Tao F (2021a) Wheat yield predictions at a county and field scale with deep learning, machine learning, and Google Earth engine. Eur J Agron 123:126204. https://doi.org/10.1016/j.eja.2020.126204

Cao J, Zhang Z, Tao F, Zhang L, Luo Y, Zhang J, Han J, Xie J (2021b) Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches. Agric for Meteorol 297:108275. https://doi.org/10.1016/j.agrformet.2020.108275

Cedric LS, Adoni WYH, Aworka R, Zoueu JT, Mutombo FK, Krichen M, Kimpolo CLM (2022) Crops yield prediction based on machine learning models: case of West African countries. Smart Agric Technol 2:100049. https://doi.org/10.1016/j.atech.2022.100049

Chandraprabha, M., & Dhanaraj, R. K. (2020). Machine learning based pedantic analysis of predictive algorithms in crop yield management. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 1340–1345. https://doi.org/10.1109/ICECA49313.2020.9297544

Crop Yield Prediction Dataset. (n.d.). . Retrieved April 22, 2024, from https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset

Darwin B, Dharmaraj P, Prince S, Popescu DE, Hemanth DJ (2021) Recognition of bloom/yield in crop images using deep learning models for smart agriculture: a review. Agronomy 11(4):4. https://doi.org/10.3390/agronomy11040646

Dubois A, Teytaud F, Verel S (2021) Short term soil moisture forecasts for potato crop farming: a machine learning approach. Comput Electron Agric 180:105902. https://doi.org/10.1016/j.compag.2020.105902

Durai SKS, Shamili MD (2022) Smart farming using machine learning and deep learning techniques. Decis Anal J 3:100041. https://doi.org/10.1016/j.dajour.2022.100041

Elavarasan D, Vincent PMD (2020) Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access 8:86886–86901. https://doi.org/10.1109/ACCESS.2020.2992480

Fan J, Bai J, Li Z, Ortiz-Bobea A, Gomes CP (2022) A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction. Proc AAAI Confer Artif Intell 36(11):11. https://doi.org/10.1609/aaai.v36i11.21444

Jayne TS, Sanchez PA (2021) Agricultural productivity must improve in sub-Saharan Africa. Science 372(6546):1045–1047. https://doi.org/10.1126/science.abf5413

Jin X-B, Yang N-X, Wang X-Y, Bai Y-T, Su T-L, Kong J-L (2020) Hybrid deep learning predictor for smart agriculture sensing based on empirical mode decomposition and gated recurrent unit group model. Sensors 20(5):5. https://doi.org/10.3390/s20051334

Kang Y, Ozdogan M, Zhu X, Ye Z, Hain C, Anderson M (2020) Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. Environ Res Lett 15(6):064005. https://doi.org/10.1088/1748-9326/ab7df9

Mateo-Sanchis A, Adsuara JE, Piles M, Munoz-Marí J, Perez-Suay A, Camps-Valls G (2023) Interpretable long short-term memory networks for crop yield estimation. IEEE Geosci Remote Sens Lett 20:1–5. https://doi.org/10.1109/LGRS.2023.3244064

Mishra P, Mohamad Alshaib B, Kuamri B, Tiwari S, Singh AP, Yadav S, Sharma D, Kumari P (2023). Forecasting potato production in major South Asian countries: a comparative study of machine learning and time series models. Potato Res https://doi.org/10.1007/s11540-023-09683-z

Nevavuori P, Narra N, Linna P, Lipping T (2020) Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models. Remote Sens 12(23):23. https://doi.org/10.3390/rs12234000

Noorunnahar M, Chowdhury AH, Mila FA (2023) A tree based eXtreme gradient boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. PLoS ONE 18(3):e0283452. https://doi.org/10.1371/journal.pone.0283452

Ortiz-Bobea A, Ault TR, Carrillo CM, Chambers RG, Lobell DB (2021) Anthropogenic climate change has slowed global agricultural productivity growth. Nat Clim Chang 11(4):306–312. https://doi.org/10.1038/s41558-021-01000-1

Pandey A, Mishra A (2017) Application of artificial neural networks in yield prediction of potato crop. Russ Agric Sci 43(3):266–272. https://doi.org/10.3103/S1068367417030028

Paudel D, Boogaard H, de Wit A, Janssen S, Osinga S, Pylianidis C, Athanasiadis IN (2021) Machine learning for large-scale crop yield forecasting. Agric Syst 187:103016. https://doi.org/10.1016/j.agsy.2020.103016

Paudel D, Boogaard H, de Wit A, van der Velde M, Claverie M, Nisini L, Janssen S, Osinga S, Athanasiadis IN (2022) Machine learning for regional crop yield forecasting in Europe. Field Crop Res 276:108377. https://doi.org/10.1016/j.fcr.2021.108377

Prasad Patnaik P, Padhy N (2023) An approach for potato yield prediction using machine learning regression algorithms. In: Kumar R, Pattnaik PK, Tavares JMRS (eds.) Next Generation of Internet of Things. Springer Nature, pp 327–336. https://doi.org/10.1007/978-981-19-1412-6_27

Qin X, Luo Y, Tang N, Li G (2020) Making data visualization more efficient and effective: a survey. VLDB J 29(1):93–117. https://doi.org/10.1007/s00778-019-00588-3

Rashid M, Bari BS, Yusup Y, Kamaruddin MA, 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

Schwalbert RA, Amado T, Corassa G, Pott LP, Prasad PVV, Ciampitti IA (2020) Satellite-based soybean yield forecast: integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agric for Meteorol 284:107886. https://doi.org/10.1016/j.agrformet.2019.107886

Shahhosseini M, Hu G, Huber I, Archontoulis SV (2021) Coupling machine learning and crop modeling improves crop yield prediction in the US corn belt. Sci Rep 11(1):1606. https://doi.org/10.1038/s41598-020-80820-1

Shetty, S. A., Padmashree, T., Sagar, B. M., & Cauvery, N. K. (2021). Performance analysis on machine learning algorithms with deep learning model for crop yield prediction. In I. Jeena Jacob, S. Kolandapalayam Shanmugam, S. Piramuthu, & P. Falkowski-Gilski (Eds.), Data Intelligence and Cognitive Informatics (pp. 739–750). Springer. https://doi.org/10.1007/978-981-15-8530-2_58

Shook J, Gangopadhyay T, Wu L, Ganapathysubramanian B, Sarkar S, Singh AK (2021) Crop yield prediction integrating genotype and weather variables using deep learning. PLoS ONE 16(6):e0252402. https://doi.org/10.1371/journal.pone.0252402

Sujatha R, Chatterjee JM, Jhanjhi N, Brohi SN (2021) Performance of deep learning vs machine learning in plant leaf disease detection. Microprocess Microsyst 80:103615. https://doi.org/10.1016/j.micpro.2020.103615

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

Wani JA, Sharma S, Muzamil M, Ahmed S, Sharma S, Singh S (2022) Machine learning and deep learning based computational techniques in automatic agricultural diseases detection: methodologies, applications, and challenges. Arch Comput Methods Eng 29(1):641–677. https://doi.org/10.1007/s11831-021-09588-5

Wolanin A, Mateo-García G, Camps-Valls G, Gómez-Chova L, Meroni M, Duveiller G, Liangzhi Y, Guanter L (2020) Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. Environ Res Lett 15(2):024019. https://doi.org/10.1088/1748-9326/ab68ac

Zaki AM, Khodadadi N, Lim WH, Towfek SK (2023) Predictive analytics and machine learning in direct marketing for anticipating bank term deposit subscriptions. Am J Bus Oper Res 11(1):79–88. https://doi.org/10.54216/AJBOR.110110

Zaki AM, Abdelhamid AA, Ibrahim A, Eid MM, El-Kenawy E-SM (2023) Enhancing K-nearest neighbors algorithm in wireless sensor networks through stochastic fractal search and particle swarm optimization. J Cybersecur Inf Manag 13(1):76–84. https://doi.org/10.54216/JCIM.130108

Authors and Affiliations

El-Sayed M. El-Kenawy[1] · Amel Ali Alhussan[2] · Nima Khodadadi[3] · Seyedali Mirjalili[4] · Marwa M. Eid[1,5]

✉ Nima Khodadadi
nima.khodadadi@miami.edu

[1] Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology, Mansoura 35111, Egypt

[2] Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, 11671 Riyadh, Saudi Arabia

[3] Department of Civil and Architectural Engineering, University of Miami, Coral Gables, FL, USA

[4] Centre for Artificial Intelligence Research and Optimisation, Torrens University Australia, Brisbane 4006, Australia

[5] Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura, Egypt