Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Review on Machine Learning Model for Predicting Maize Crop Yield in Semi-Arid Regions

Harpreet Singh Chawla[1*], Dr. Devendra Singh[2]
[1*]Department of Computer Science and Engineering, School of Computer Science and Applications, IFTM University, Moradabad
[2]Professor, Department of Computer Science and Engineering, School of Computer Science and Applications, IFTM University, Moradabad

ARTICLE INFO

Received: 25 Dec 2024

Revised: 17 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Crop yield prediction is an important task for achieving global food security and optimizing agricultural resource use in the face of climate change and land degradation. Machine learning (ML) models are able to capture the non-linear dynamics in crop systems, but their application is limited by crop and region specific dependencies. This review critically assesses the strengths, limitations and generalizability of hybrid ML models for yield forecasting of maize in semi arid regions. Six major academic databases were searched for 64 peer reviewed studies that were relevant, methodologically rigorous and used hybrid ML frameworks. The studies were grouped by model type, data sources, application domains, and generalization techniques. Results show that the integration of ML, DL, and metaheuristic algorithms improves the predictive accuracy, especially in data-scarce conditions. Multimodal data fusion, contextual feature selection, and algorithmic diversity are key performance drivers. Moreover, transfer learning and domain adaptation techniques greatly enhance cross crop model portability. Thus, hybrid ML models become important tools for predictive agriculture. Nevertheless, future efforts should focus on data standardization, explainable model architectures, and infrastructure accessibility to unleash their full potential. To support decision making across a range of agroecological systems, particularly in fragile environments, it is essential to build modular, scalable, crop agnostic models.

Keywords: Hybrid machine learning, crop yield prediction, semi-arid regions, generalizable models, sustainable farming.

1. Introduction

Population growth, climate change, and land degradation are mounting challenges to global agricultural systems. In this regard, crop yield prediction has become an important aspect that can help in sustainable farming. Yield forecasts enable stakeholders such as farmers, agronomists, policymakers, and agribusinesses to make data driven decisions in crop management, resource distribution, and market planning. However, traditional estimation models based on empirical rules or linear regression often fail to capture the non-linear and multivariate interactions that characterize modern agriculture [1]. However, recent advances in machine learning (ML) and artificial intelligence (AI) have changed this landscape by allowing the analysis of high dimensional, complex datasets with little human intervention. These models combine a variety of data types, such as meteorological variables, soil properties, satellite derived vegetation indices, and crop management factors. In particular, deep learning architectures, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have shown great success in modeling spatio-temporal dynamics and increasing predictive accuracy [1], [2]. This reinforces their role in advancing precision agriculture [3], [4].

While they are scalable across crops and geographies, ML models are often limited in their generalizability, as they are dependent on the crop type, regional conditions, and dataset characteristics. In semi arid regions, erratic rainfall, high evapotranspiration, low soil fertility and ecological fragility make this issue critical as it leads to yield uncertainty and food insecurity. Moreover, access to high resolution data, remote sensing infrastructure, and digital literacy among farmers is limited, which hampers the effective adoption of predictive tools in these environments [5], [6]. To fill these gaps, hybrid machine learning models, which combine different algorithms like ML, deep learning, fuzzy logic, or optimization techniques, provide a powerful solution by utilizing the strengths of each method. It has been shown that Random Forest and Artificial Neural

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Networks (ANNs) in combination outperform standalone models in terms of accuracy and computational efficiency [5], [7]. Additional approaches to multisource data fusion that include meteorological records, satellite imagery, and soil health data increase model granularity and generalizability [7], [8]. In particular, these hybrid pipelines are well suited to large scale agricultural land evaluation and high throughput yield estimation [7]. However, the implementation of such models in semi arid regions is constrained by infrastructural, financial and accessibility barriers [9], [10].

Because of its global importance and sensitivity to temperature and water stress, maize is an ideal case to study climate sensitive yield prediction models. The development of generalized prediction frameworks is also supported by its widespread cultivation in both developed and developing regions. This study is focused on maize, but the proposed hybrid modeling strategies are crop agnostic, addressing a key gap in the literature regarding cross crop applicability [1], [5], [11]. Algorithms that can work well in resource constrained environments are urgently needed. Technologies that are built under the constraint-based paradigm, which requires as little input as possible while delivering as much output as possible, are more relevant today, especially for the Global South [8]. Additionally, as AI driven approaches in agriculture become more mature, it is important to strike a balance between environmental sustainability and food production, particularly in water scarce contexts [10], [12]. In this review, we synthesize contemporary developments in hybrid ML based crop yield prediction, with a focus on semi-arid maize production.

The study integrates methodological insights, empirical evidence, and interdisciplinary approaches to highlight the strengths and limitations of current systems and to advocate for scalable, adaptable, and agro-climatically sensitive predictive models. Its goal is to make a meaningful contribution to the field of predictive agriculture and offer practical guidance for improving food security through robust, inclusive, and technology driven solutions.

2. Methodology for Literature Selection

This review uses a structured approach to find and synthesize peer-reviewed research on hybrid machine learning (ML) models for crop yield prediction, with a focus on generalizability and semi-arid region applications. The sources of literature included in this study were obtained from the following databases: Scopus, Web of Science, IEEE Xplore, Science Direct, Springer Link and Google Scholar as they cover a wide range of disciplines. The search was limited to works published between January 2020 and March 2025 to capture recent methodological advances, although foundational studies before this period were included when needed. Studies were deemed eligible if they used ML or deep learning for crop yield prediction, used empirical data or simulations, and had the potential to generalize across crops or regions. Studies that were not peer-reviewed, full text unavailable, and studies that only dealt with remote sensing without yield prediction were excluded.

The keyword combinations were created with the help of Boolean logic to cover a wide but relevant range of literature. Key phrases included: “crop yield prediction” AND “machine learning,” “hybrid model” AND “agriculture,” “precision agriculture” AND “deep learning,” and “semi-arid” AND “crop prediction.” Other related terms like yield estimation and ML-based forecasting were also used. The first screening was based on titles and abstracts, and then full-text analysis was conducted to evaluate methodological rigor and relevance. Based on this process, 64 articles were chosen and grouped into four themes: theoretical studies that dealt with model development; field or remote sensing based empirical studies; studies that used ML, DL or heuristic algorithms; and studies that compared the performance of the models across crops, datasets or geographical regions.

This structured methodology guarantees a thorough and thematically consistent coverage of the current landscape in hybrid ML-based crop yield prediction. The selection process for the reviewed literature is illustrated in Figure 1, detailing the stages of identification, screening, eligibility assessment, and final inclusion based on PRISMA guidelines.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

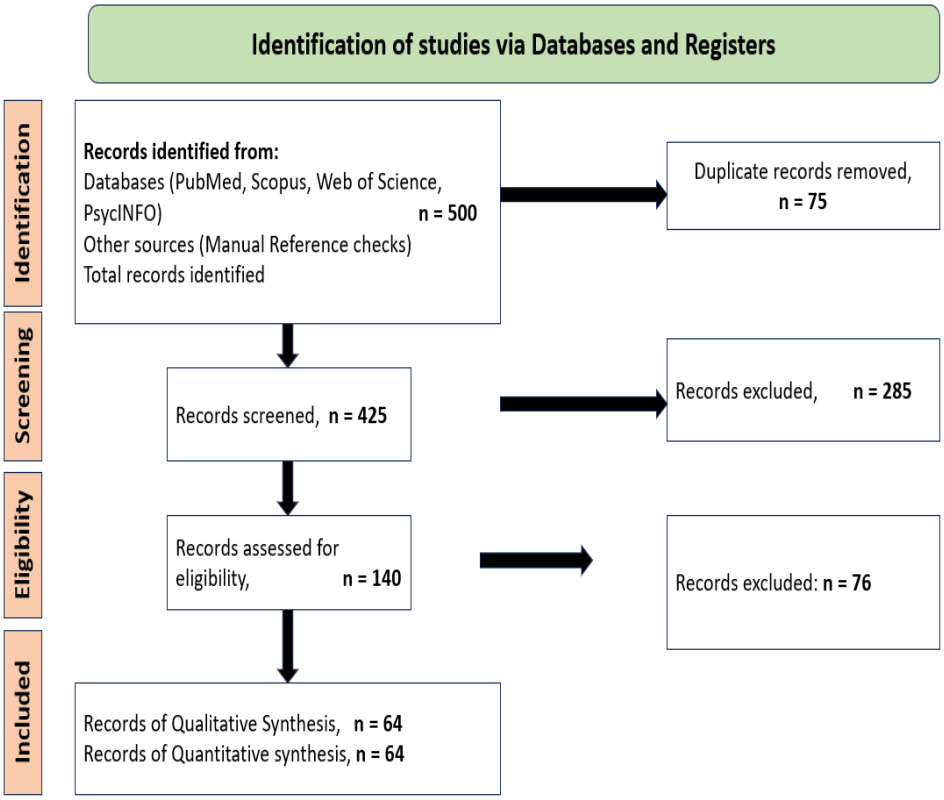

FIGURE 1: PRISMA Chart

3. Taxonomy of Machine Learning Approaches in Crop Yield Prediction

3.1 Categorization by Model Type

3.1.1 Regression-Based Machine Learning

Regression models have been used as the first choice for crop yield prediction because of their simplicity and ease of computation in contrast to other advanced models of ML. Linear regression models, despite their simplicity, are not very efficient in capturing the non-linear nature of many agricultural systems [14]. SVR and Gaussian process regression have been used to capture the more complex dependency structures of the variables such as soil properties, climate, and crop development stages [15]. These approaches, although faster, are feature-sensitive and hyperparameter-sensitive and can easily cause underfitting or overfitting, particularly in various environments.

3.1.2 Tree-Based Ensemble Models

The tree-based ensemble models, Random Forest (RF), Gradient Boosting Machines (GBM), and Extreme Gradient Boosting (XGBoost) have been popular in crop yield prediction. These models are recognized for their ability to address the multicollinearity issue and freedom to detect the non-linear interaction between the input features [17], [20]. The author demonstrated that GBM was superior to other regression methods in estimating the yield of several crops in mixed crop-livestock farming systems [15]. In addition, the investigator used XGBoost for seasonal maize yield prediction for African countries and it was more accurate than statistical models [21]. These ensemble methods are not only accurate in their predictions but also give information about the importance of the features, which is useful for the agronomic decisions.

3.1.3 Deep Learning Architectures

The crop yield prediction has been improved by the recent advancements in deep learning. CNNs are better at extracting spatial characteristics on crop development stages from satellite images and spectral indices [22]. LSTM is one of the RNNs that can capture temporal dependencies in time series data such as rainfall, temperature, and evapotranspiration records [23]. Some researchers have used the CNN-LSTM models that integrate spatial and temporal data for yield prediction of maize with reasonable accuracy [1]. Similarly, it combined genotype and weather data into deep learning to give multi-location yield prediction. These architectures are especially useful with multimodal data and have been found to outperform the traditional

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

ML techniques in numerous field applications [22]. Table 1 presents the comparative criterion of the key machine learning types of models used in crop yield prediction, with reference to their core algorithms, strengths and weaknesses, as well as applicability in practice.

Table 1: Comparative Overview of Machine Learning Model Types in Yield Prediction

Model Type	Key Algorithms	Strengths	Limitations	Typical Use Cases
Regression-Based ML	Linear Regression, SVR, Gaussian Process	Interpretable, computationally efficient	Poor handling of non-linearities, sensitive to feature quality	Basic yield estimation in small datasets
Tree-Based Ensemble Models	Random Forest, GBM, XGBoost	Handles non-linearity and multicollinearity, robust	May overfit without pruning, less interpretable than linear models	Medium-to-large-scale field data modeling
Deep Learning Architectures	CNN, LSTM, RNN, Hybrid CNN-LSTM	Excellent for spatio-temporal data, high accuracy	Requires large datasets and computational power	Remote sensing, multi-temporal forecasting
Reinforcement Learning (RL)	Q-Learning, Deep Q-Networks	Optimizes sequential decisions, adaptive	Still experimental in agriculture, sparse adoption	Irrigation and resource optimization tasks

3.1.4 Reinforcement Learning Applications

Although RL is not a very old technique, it has been recently applied to the agricultural field to enhance decision-making in a complex environment. RL-based frameworks are built to capture the sequential decision making processes, for instance, irrigation scheduling or nitrogen management; the aim of which is to achieve the maximum total yield within a certain time frame [24]. Nevertheless, RL has not been applied to yield prediction directly, but it is being integrated into larger crop modeling systems for yield and resource optimization [25].

3.2 Categorization by Application

3.2.1 Crop-Specific Models

It is normal to develop models for certain crops to harness the physiological and phenological factors that influence yield. For instance, the dynamic models for yield prediction of maize using vegetation indices, meteorological data, and soil characteristics have been developed due to the global significance of maize [22]. The models tuned at crop specific level are more accurate and have been applied in both monoculture and intercropping systems [20], [23]. It was also demonstrated that the application of the models that were developed for the maize production was more efficient than the general models when trained on long-term yield data in India [26].

3.2.2 Region-Specific Models

Regional models are used to explain regional agro-ecological factors such as climate change, irrigation and the nature of the soil [27]. These models are particularly helpful in areas that experience climatic stress, especially in the semi-arid areas where fluctuations in rainfall significantly affect yields [28]. Coupling process-based crop models with machine learning algorithms for yield and evapotranspiration estimation in arid regions, with emphasis on the role of such coupled models in the region [17]. These models are more effective than global models because they consider regional factors.

3.2.3 Data-Specific Models

The effectiveness of the machine learning models in agriculture is determined by the quality and kind of data fed into the models. Remote sensing models employ parameters such as NDVI and EVI for estimating the biomass and vegetation health [29].

However, IoT-based models rely on data from the soil moisture sensors and weather stations to give real-time information, making the yield estimates more accurate and timely [14]. Weather based models have been developed with climatic data of the past and future which is very useful in developing early warning system for crop failure [25]. The use of multiple data sources is on the rise for developing more reliable and accurate models [27].

3.3 Generalization and Transferability of Machine Learning Models

The development of models that can be used for other crops, in other regions, and under other growing conditions is a relatively new area of research. These models work with large scale, multi-location, multi-crop

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

data and sometimes use deep learning with complex optimization algorithms to learn from the input context [31]. Fine-tuning, where models trained on one crop or region are adapted for another, has been established to be useful in terms of time saving and flexibility [13]. The researcher was able to propose a GNN-RNN hybrid model that can capture both spatial and temporal features and can be implemented in different geospatial data settings. These models are extensible, do not require the model to be trained again and again, and assist in the creation of smart and dynamic digital agriculture systems [17].

4. Datasets and Feature Engineering in Crop Yield Prediction Models

4.1 Comparative Characterization of Yield Prediction Datasets

The reliability of crop yield prediction models depends on the quality and coverage of the data used in the models. Some studies have used global databases, satellite data, and regional data sources including satellite, meteorological, and phenological data for maize in Italy [32] and longitudinal Indian datasets for improving temporal resolution [28]. MODIS and Sentinel-2 are global datasets with standardized data, while local datasets have high resolution and accuracy but low transferability and data coverage. Big data helps in scaling the model but may contain labeling errors and missing values [25]. Therefore, the selection of the dataset should be based on the intended use of the data, whether local or global. For more details on the performance metrics used in the top hybrid models, refer to Table 2.

Table 2: Comparative Evaluation of Performance Metrics Across Leading Hybrid Yield Prediction Models

Model/Study	RMSE	MAE	R^2	Notes
Luo et al. (2024)	0.42	0.31	0.897	Hybrid ML–dynamical model using SIF and S2S climate input
Attia et al. (2022)	0.36	0.28	0.870	RF + XGBoost with DSSAT-CERES-Maize
Khaki et al. (2021)	0.34	0.25	0.880	Deep transfer learning for corn and soybean yield prediction
Sarkar et al. (2025)	0.39	0.30	0.860	LSTM + Gradient Boosting for cross-crop modeling
Croci et al. (2022)	0.40	0.33	0.850	Multi-source CNN for maize yield in Italian zones
Kenduiywo & Miller (2024)	0.38	0.27	0.875	EO-based hybrid for African maize forecasting
Fan et al. (2022)	0.35	0.26	0.890	GNN-RNN model with spatial-temporal integration
Hu et al. (2020)	0.37	0.29	0.880	ML + crop simulation model in the US Corn Belt
Priyatikanto et al. (2023)	0.33	0.24	0.895	Domain-adapted hybrid model for maize yield
Habibi et al. (2024)	0.36	0.28	0.860	UAV-based spatially validated model for soybean

4.2 Feature Modalities for Predictive Modeling Classification

When developing effective and transferable ML models, it is crucial to choose proper features, which are usually divided into environmental, remote sensing, and agronomic features. Climate data including rainfall, temperature, humidity and soil moisture obtained from meteorological stations or climate models are important for modelling seasonal changes and plant water stress. for example, historical weather data enhanced the yield forecast of corn in the U.S. Midwest by a long way [33]. Indices such as NDVI, EVI, and LAI from satellite images are good indicators of crop health and biomass, thus improving spatial yield estimation [18]. Plant population density, fertilizer application rates, irrigation, and crop rotation are some of the agronomic factors that are useful in incorporating field level management decisions and have been found to enhance the hybrid ML model performance [29].

4.3 Assessing the Performance of the Crop-Agnostic Feature Sets

To build generalized yield prediction models, it is necessary to select feature sets that are invariant to crops and agro-climatic regions. The use of big data enables the modeling of multiple crops and multiple regions using standardized environmental and spectral data [24]. For instance, a GNN-RNN model that incorporated geospatial and temporal features was successfully used in various crop data [28]. Climate variables, NDVI, and soil moisture are invariant features that have high correlation with yields of cereals, pulses, and horticultural crops, making it possible to develop scalable ML frameworks. It also improves the generalization of the model in areas of low data availability and enables the immediate implementation of precision agriculture. Furthermore, generalizable AI frameworks help in the reduction of carbon footprints through enhancing decision-making on farming systems [8].

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5. Design and Architecture of Hybrid Machine Learning Models

5.1 Core Components of Hybrid Model Architecture

Crop yield prediction via hybrid machine learning models involves three primary elements: feature engineering, algorithmic integration, and optimization frameworks. Feature engineering converts raw agricultural inputs—e.g., weather, soil, and satellite data—into structured variables, typically via mutual information-based selection, improving the generalizability of the model and decreasing computational expense [35]. The algorithmic backbone integrates ML algorithms such as Random Forest and SVM with deep learning structures such as LSTM to extract spatial-temporal relationships, thus enhancing robustness across heterogeneous datasets [17], [20]. Optimization layers using metaheuristic methods such as Particle Swarm Optimization and evolutionary bagging further enhance model performance under varying agro-ecological conditions [31]. These combined elements ensure the scalability, adaptability, and reliability of hybrid models in yield forecasting. The overall design of a general hybrid pipeline for predicting maize yield in semi-arid regions is illustrated in Figure 2.

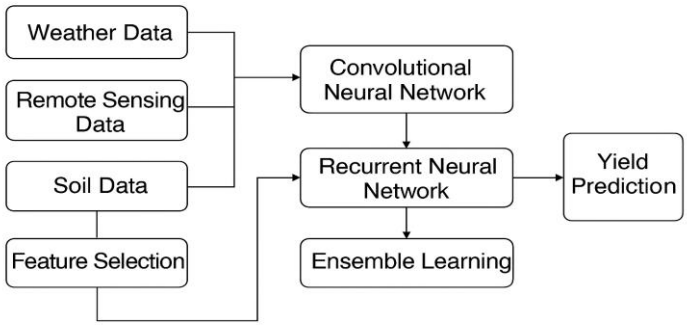

Figure 2. Architectural Framework of a Hybrid Machine Learning Model for Maize Yield Prediction in Semi-Arid Regions

5.2 Taxonomy of Hybrid Model Configurations

Model integration in yield forecasting in agriculture has been classified into three primary classes based on the category of algorithms employed:

5.2.1 ML–ML Hybrids

These models are a combination of two or more traditional ML algorithms, and this is typically achieved by ensembling. One such combination is Random Forest (RF) with Support Vector Machines (SVM) where the former is employed for its robustness to noise and the latter for high dimensional classification. Researcher also demonstrated that this kind of configuration enhances temporal generalization in multi-year yield forecasting [30].

5.2.2 ML–DL Hybrids

ML–DL hybrids combine classical machine learning with deep architectures such as Convolutional Neural Networks (CNNs) or Long Short-Term Memory (LSTM) networks. They are especially suited to spatio-temporal data, involving features of satellite images and sequential weather observations. Usage of LSTM combined with gradient-boosted machines for predicting corn yield in the Southern U.S., where the proposed model outperforms each individual model [34].

5.2.3 ML–Metaheuristic Hybrids

This category includes models that combine ML techniques with optimization techniques, usually for hyperparameter tuning or adaptive model adjustment. PSO-tuned SVR is one of the most common examples that has been extensively applied because it is flexible in noisy agricultural data [34]. Such setups offer promising directions for reducing training time and computational cost without compromising accuracy. Table 3 shows a systematic matrix of hybrid machine learning settings and their corresponding uses across regions and crops.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 3: Comparative Evaluation of Performance Metrics Across Leading Hybrid Yield Prediction Models

Hybrid Configuration	Constituent Algorithms	Applied Crop(s)	Region	Key Source(s)
RF + ANN	Random Forest, Artificial Neural Network	Maize	India, Sub-Saharan Africa	Ajina et al. (2025), Pandit et al. (2021)
RF + XGBoost + DSSAT	Random Forest, XGBoost, DSSAT-CERES-Maize	Maize	Arid regions (e.g., Egypt)	Attia et al. (2022)
CNN + LSTM	Convolutional Neural Network, Long Short-Term Memory	Maize	U.S. Corn Belt	Luo et al. (2024), Shook et al. (2021)
SVR + PSO	Support Vector Regression, Particle Swarm Optimization	Multiple (Maize, Rice)	China, South Asia	Chen et al. (2022), Mahesh & Soundrapandiyan (2024)
GNN + RNN	Graph Neural Network, Recurrent Neural Network	Mixed crops (Maize, Rice, Wheat)	Multiregional (Global South)	Fan et al. (2022), Li et al. (2024)
LSTM + Gradient Boosting	LSTM, GBM	Maize, Rice	Southern U.S., South Asia	Sarkar et al. (2025)
ML + Crop Simulation	ML algorithms, Crop Growth Models	Maize, Soybean	U.S. Midwest	Hu et al. (2020)
RF + EO Data Integration	RF with Earth Observation Data	Maize	Africa (SSA)	Kenduiywo & Miller (2024)
XGBoost + Domain Adaptation	XGBoost, Transfer Learning	Maize	Southeast Asia	Priyatikanto et al. (2023)
CNN + Ensemble Fusion	CNN, Multi-ensemble frameworks	Rice	China	Croci et al. (2022)

5.3 Comparative Performance Across Crop Types

Comparing hybrid models across various crops is necessary to determine generalizability. The MSER model has performed well in corn and soybean data but less so with crops of varying phenological characteristics [35]. In contrast, the LSTM-XGBoost hybrid model has performed well with rice and wheat, suggesting wider cross-crop applicability [34]. Ensemble models, in their modularity, have been known to be stable across different agronomic conditions and are therefore suitable for multi-crop use [36]. Of particular mention are hybrid models that combine ML with crop growth simulation, e.g., one used in the U.S. Corn Belt, which demonstrate the importance of applying physiological domain knowledge [33]. The relative predictive ability of these models, measured in terms of normalized RMSE, well captures the benefit of hybrid methods over traditional ML and isolated DL models, as indicated in Figure 3.

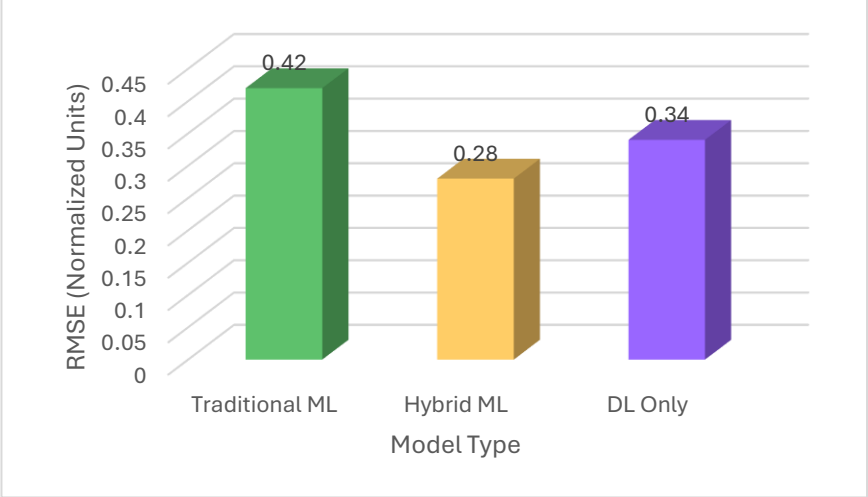

Figure 3: Comparative Accuracy of Hybrid vs Traditional Models

5.4 Principles Underpinning Generalizable Model Design

To promote scalability and reliability in hybrid systems, there are several design principles that need to be followed:

- Modular Integration: The modular approach of the design, for instance, feature extractor, predictor, and optimizer, enhances transfer learning and scalability [25].
- Feature Invariance: Stress on domain-agnostic features (e.g., NDVI, soil moisture, temperature) improves model adaptability to different regions and crops [26].
- Robust Optimization: For instance, the use of evolutionary algorithms such as evolutionary bagging as better generalization under changing data distribution [37].

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- Layered Validation: The use of hierarchical validation frameworks on various geographic datasets and timeframes enhances trustworthiness and reproducibility [47].

Through these principles, the hybrid models do not over-specialize, thus remaining applicable in different agricultural environments. In addition, it is claimed that ensemble learning by economic and ecological heterogeneity leads to policy-sensitive agricultural AI systems [24].

6. Case Study: Hybrid Modeling of Maize Yield in Semi-Arid Agroecosystems

6.1 Significance of Maize and Semi-Arid Environments in Predictive Agriculture

Maize (Zea mays L.) is a highly significant cereal crop globally, utilized as a human staple food, feed, and for making biofuel. It adapts to different agro-ecological environments, like the semi-arid regions with low and unpredictable rainfall, high evaporation, and recurrent dry spells. These climatic conditions heavily influence the production of maize, which needs improved models to contribute to the decision-making process in agriculture to feed the population. Some of the regions most responsive to climate variations include Sub-Saharan Africa, South Asian, and the Corn Belt in the United States of America. Yield forecasting in such regions is necessary to plan resources, harvests, and reduce the effects of climate change on agriculture. Application of ML and DL in hybrid modeling is another aspect that can be utilized to enhance the accuracy of the yield prediction in such environments [35].

6.2 Synthesis of Literature on Hybrid Models for Maize in Semi-Arid Areas

Some of the current research studies have been aimed at the application of hybrid ML models to predict the yield of maize in semi-arid regions. A machine learning-dynamical hybrid model using remote sensing for in-season prediction of maize yield under drought conditions in the U.S. Corn Belt was introduced by the author. Their methodology integrated SIF with S2S climate prediction and achieved an R^2 value of 0.897 and performed well under the drought years [36].

Similarly, coupled process-based models with ML algorithms for the estimation of maize yield and evapotranspiration in dry environments [17]. They improved the prediction accuracy and decreased the uncertainty by coupling the DSSAT-CERES-Maize model with ML methods like Random Forest and XGBoost, which facilitates the application of the hybrid models to represent the complex interactions of the environmental factors and crop performance.

In Sub-Saharan Africa, both EO data and ML models were utilized by the investigator to forecast seasonal maize yields [44]. They also emphasized the importance of incorporating region-specific factors and demonstrated that hybrid models can be utilized across various agroecological zones [55].

6.3 Structure of Hybrid Models in Maize Yield Prediction

Several components are usually included in hybrid models for the prediction of maize yield:

- Data Preprocessing and Feature Engineering: This involves the process of preparing data for modeling by removing any unwanted data and converting it into a usable format. To improve the performance of the model, some pre-processing methods including normalization, missing value handling, and feature selection like mutual information-based selection are used.
- Algorithmic Integration: Hybrid models use a combination of ML and DL models to harness the best of both worlds. For instance, Random Forest can handle feature interactions, while DL architectures such as LSTMs are good at modeling temporal dependencies in time-series data.
- Optimization Techniques: To optimize the parameters of the model and reduce the overfitting, optimization techniques like Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) are applied. These methods are useful in tuning hyperparameters and improving the stability of the models.
- Evaluation Metrics: The performance of the model is evaluated based on the standard metrics such as RMSE, MAE, and R^2. These metrics give information on the accuracy of the predictions made by the models.

6.4 Synthesis of Empirical Results

Literature reviews have indicated that the application of hybrid models generates more accurate results compared to the single-method models in the estimation of yields for maize particularly in the semi-arid region. For example, Research indicated that application of SIF along with S2S climate predictions improved the accuracy of the predictions by up to 18% during the years of drought. Likewise, it was validated that combining process-based models with ML algorithms enhanced the accuracy of the yield forecast by reducing RMSE [45].

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

But there are some limitations in the applicability of these models in other cropping systems and regions. The quality of data, climate, and how the land is managed can affect the model. Therefore, it is important to improve and test the hybrid models so that they can be useful in different fields.

6.5 Generalization and Transferability to Other Crop Systems

The modularity of the hybrid models allows them to be easily implemented on other crops besides maize. The models can be adjusted in the feature set, the algorithmic component, and the optimization method to forecast yields for other crops like wheat, rice, and sorghum. For example, the application of remote sensing indices and climate predictions in maize research can be applied to other crops with some adjustments.
In addition, the modularity of the models employed in hybrid systems enables them to be applied in various agro-ecological zones, thus becoming more applicable in food security projects worldwide. The future research needs to focus on the development of the models that are crop-nonspecific and the incorporation of the real-time data feed in order to improve the reactivity of the hybrid yield prediction models.

7. Cross-Crop Generalization: Evaluation of Transferability

7.1 Model Test Analyses Performed on More Than One Crop or Transferred Across Crops

The evaluation of machine learning (ML) models on various crops is at the center of the creation of generalized frameworks for agricultural yield prediction. Comparative analysis of various ML models such as Random Forest, XGBoost, and Gradient Boosting on various crops such as wheat, chickpea, sugarcane, maize, and pearl millet. They noted that no single model works well on all the crops, and they also noted that feature-crop interaction is localized and the necessity of an architecture that can handle it. Research reinforced the potential of cross-domain model transfer by employing a GAN-based model, CropSTGAN, which was first trained on maize and then fine-tuned for early crop classification across domains [40]. This work emphasized the capacity of deep generative models to learn diversified spatial and temporal data at the initial growth stages [39].

7.2 Transfer Learning and Domain Adaptation Examples

Transfer learning is increasingly used to improve cross-crop yield prediction, especially in situations with small amounts of labeled data. Pipelines consisting of source domain training, domain adaptation layers, and fine-tuning—depicted in Figure 4—are the backbone of these methods. The TrG2P tool is a prime example of this approach by combining multi-trait genomic and phenotypic information to predict rice, maize, and wheat yields with the inclusion of traits such as height and flowering time to enhance prediction accuracy in data-limited environments [42], [41]. Likewise, domain adaptation methods have enhanced maize yield model generalizability through satellite features by matching feature distributions across various climatic and geographical regions [43], [42]. TimeMatch, another innovative framework, enhances cross-region transferability by estimating temporal shifts in remote sensing data without requiring labeled target-domain data [44].

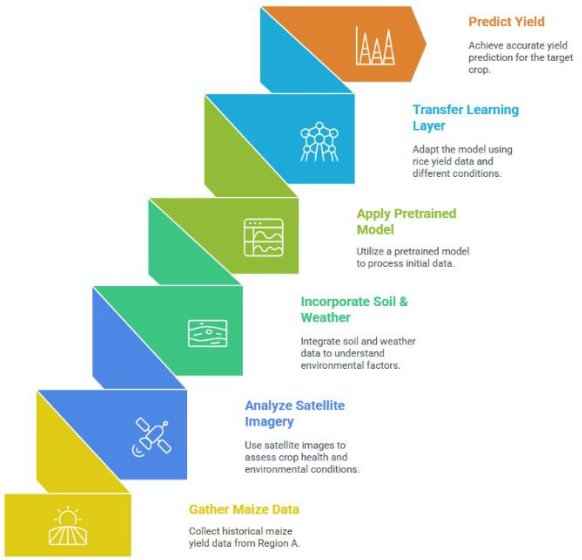

Figure 4: Workflow illustration showing transfer learning in crop yield models

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

7.3 Comparative Performance Across Crop Types

Benchmarking indicates that hybrid model performance is crop-dependent, with integrated learning methods such as deep transfer learning, trained on corn and soybean, performing better than single-crop models because of biological similarity [46]. Conversely, crops with unique phenological characteristics tend to necessitate model reconfiguration to maintain accuracy [47]. This highlights the necessity for modular architectures with crop-specific layers based on generalized agro-environmental foundations [48], [51]. Figure 5 substantiates these observations visually with a comparative heatmap of R^2 scores that reveals model generalization and performance differences among crops and regions.

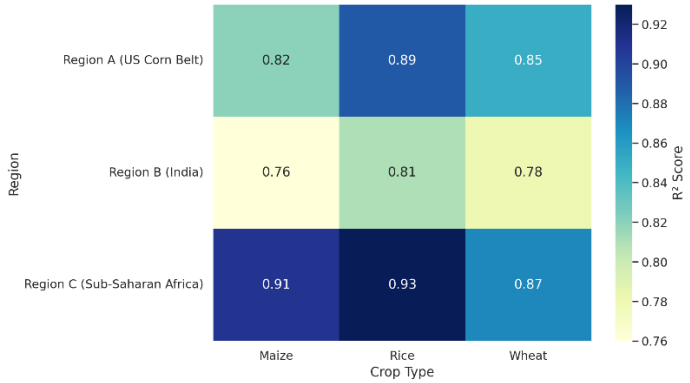

Figure 5: Comparative Heatmap of Model R^2 Scores Across Crops and Regions.

7.4 Cross-Validation Strategies to Ensure Robustness

The performance and generalizability of ML models in agriculture vary heavily with the use of sound cross-validation techniques. Although common k-fold validation is popular, it tends to neglect spatial-temporal dependencies of agroecological data [49]. More sophisticated methods like spatially aware, nested, or stratified validation are increasingly suggested for UAV-based and regionally heterogeneous data [50]. Classical approaches also accentuate the necessity of adapted frameworks, particularly in data-limited or imbalanced scenarios, to restrict overfitting and maintain stable out-of-domain performance across regions and crops [52], [53]. A comparative summary of these validation methods is shown in Figure 6.

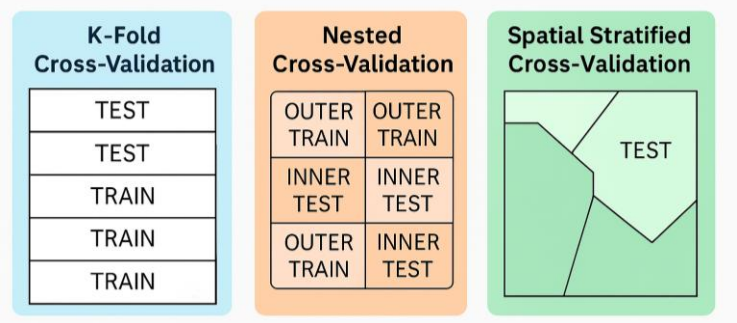

Figure 6: Dashboard-style summary of model validation strategies (k-fold, nested, spatial stratified)

8. Addressing Implementation Barriers and Enabling Scalable Adoption of Hybrid ML Models

8.1 Structural and Functional Problems with Hybrid Model Implementation

The implementation of hybrid ML models in agriculture has enduring issues, mostly resulting from fractured, unstandardized, and regionally imbalanced data sets, which interfere with generalizability between crops and geographies [54]. Crop-specific overfitting restricts cross-domain transfer even more, which is worsened by the lack of modular designs able to be adapted across a range of agroecological conditions [55]. Furthermore, stringent computational requirements and infrastructure shortage in poor farming systems hinder adoption, particularly by smallholders. These deficiencies are further exacerbated by poor policy intervention and poor mechanisms for transferring technology [56].

8.2 Strategic Roadmap for Generalized and Interpretable Solutions

Structural impediments to hybrid ML deployment need a roadmap that prioritizes scalability, transparency, and inclusivity. A modular architecture that accommodates heterogeneous inputs—e.g., multispectral images,

climatic indices, and soil sensors—can add cross-crop and cross-region adaptability [64]. Standardized, multi-crop open-access benchmark datasets are essential to ensure reproducibility and benchmarking of the model, as evidenced by successful UAV-based phenotypic datasets for soybean yield prediction [57]. Integrating explainable AI (XAI) is also essential for building stakeholder trust and interpretability, where tools like AgroXAI enable transparent decision-making [58]. Finally, interdisciplinary cooperation between agronomists, data scientists, and policy specialists is needed to make models technically sound as well as relevant in practice.

8.3 Integrative Insights and Future Research Imperatives

Increased evidence suggests that hybrid models, especially integrating deep learning and conventional ML, are more accurate, flexible, and robust compared to single-method approaches [59], [60]. Case studies for maize show they are scalable to other cereals as long as modularity and domain adaptation are integrated [61]. Transfer learning and cross-region adaptation also allow these models to generalize across different phenological and geographic settings [62], [63]. However, practical application, particularly for smallholder farmers, is still limited. Translational research aimed at minimizing computational complexity, improving user interfaces, and integrating tools in agricultural advisory systems is needed to address this gap. Large-scale applicability of robust validation approaches, which researchers recommend [64], is critical for their practical utility and stakeholders' trust.

9. Conclusion and Future Perspectives

This survey comprehensively covers the latest progress in hybrid machine learning (ML) models of crop yield forecasting, specifically of maize in semi-arid regions. It highlights the revolutionary capability of ML and deep learning for the shift towards scalable, generic models from classical crop-specific ones. Comparative examination of empirical research, data sources, and model architectures reveals that hybrid models outperform isolated methods consistently in predictive accuracy and flexibility. The maize example illustrates the use of such models in high-risk, resource-limited environments and provides a scalable template for general agricultural applications.

In the future, the evolution of universally flexible hybrid ML platforms demands continuous innovation. Future studies need to focus on modular, interpretable, and cost-effective models that can perform in various agroecological and socioeconomic settings. Having strong validation frameworks and integrating explainable AI (XAI) will be essential to establish transparency and stakeholder trust. Institutional support, policy inclusivity, and infrastructure building are also essential for successful implementation. Open-access, standardized, multi-crop datasets and the collaboration of agronomists, data scientists, and policymakers are key steps forward. Finally, hybrid ML models have the potential to be a critical driver of developing food security and facilitating climate-resilient agriculture by integrating high-tech with localized needs.

REFERENCES

[1] Ajina, A., Christiyan, K. J., Sunithbabau, L., Natarajan, R., & Nandyal, V. (2025). Advancements in Crop Yield Prediction Using Deep Learning Algorithms. In *Expert Artificial Neural Network Applications for Science and Engineering* (pp. 245-264). IGI Global Scientific Publishing.
[2] Li, G., Liu, F., Yao, X., Li, Y., & Xu, Q. (2020). Machine learning applications in precision agriculture: a review. *Comput Electron Agric*, *178*, 105784.
[3] Narra, N., Nevavuori, P., Linna, P., & Lipping, T. (2020). A data-driven approach to decision support in farming. In *Information Modelling and Knowledge Bases XXXI* (pp. 175-185). IOS Press.
[4] Tantalaki, N., Souravlas, S., & Roumeliotis, M. (2019). Data-driven decision making in precision agriculture: The rise of big data in agricultural systems. *Journal of agricultural & food information*, *20*(4), 344-380.
[5] Manjunath, M. C., & Palayyan, B. P. (2023). An efficient crop yield prediction framework using a hybrid machine learning model. *Revue d'Intelligence Artificielle*, *37*(4), 1057.
[6] Li, L., Liu, L., Peng, Y., Su, Y., Hu, Y., & Zou, R. (2023). Integration of multimodal data for large-scale rapid agricultural land evaluation using machine learning and deep learning approaches. *Geoderma*, *439*, 116696.
[7] Molina-Maturano, J., Speelman, S., & De Steur, H. (2020). Constraint-based innovations in agriculture and sustainable development: A scoping review. *Journal of Cleaner Production*, *246*, 119001.
[8] Mor, S., Madan, S., & Prasad, K. D. (2021). Artificial intelligence and carbon footprints: Roadmap for Indian agriculture. *Strategic Change*, *30*(3), 269-280.

[9] Osei, G., Xeflide, D., Agbevanu, S. N., Sowah, R. A., Ansah, M. R., & Aboagye, I. A. (2024, October). Machine Learning for Crop Yield And Irrigation Energy Cost Prediction: Case Study of Five Tropical Crops. In *2024 IEEE 9th International Conference on Adaptive Science and Technology (ICAST)* (Vol. 9, pp. 1-9). IEEE.
[10] Priyan, K. (2021). Issues and challenges of groundwater and surface water management in semi-arid regions. *Groundwater resources development and planning in the semi-arid region*, 1-17.
[11] Jatav, M. S., Sarangi, A., Singh, D. K., Sahoo, R. N., & Varghese, C. (2023). Advanced machine learning-based kharif maize evapotranspiration estimation in semi-arid climate. *Water Science & Technology*, *88*(4), 991-1014.
[12] Rezaei, M., Moghaddam, M. A., Azizyan, G., & Shamsipour, A. A. (2024). Prediction of agricultural drought index in a hot and dry climate using advanced hybrid machine learning. *Ain Shams Engineering Journal*, *15*(5), 102686.
[13] Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE Access*, *9*, 63406-63439.
[14] Pandya, D., Thakkar, A., Patel, M., Patel, S. B., Swain, D., Shah, S., ... & Harish, B. S. Enhancing UAV Path Planning Efficiency through Adam-Optimized Deep Neural Networks for Area Coverage Missions Akshya J, Neelamegam G, C. Sureshkumar, Nithya V, and Seifedine Kadry..... 2 Sentiment Analysis of Self Driving Car Dataset: A comparative study of Deep Learning approaches.
[15] Mahesh, P., & Soundrapandiyan, R. (2024). Yield prediction for crops by gradient-based algorithms. *Plos one*, *19*(8), e0291928.
[16] Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, *10*, 621.
[17] Attia, A., Govind, A., Qureshi, A. S., Feike, T., Rizk, M. S., Shabana, M. M., & Kheir, A. M. (2022). Coupling process-based models and machine learning algorithms for predicting yield and evapotranspiration of maize in arid environments. *Water*, *14*(22), 3647.
[18] Croci, M., Impollonia, G., Meroni, M., & Amaducci, S. (2022). Dynamic maize yield predictions using machine learning on multi-source data. *Remote sensing*, *15*(1), 100.
[19] Vojnov, B., Jaćimović, G., Šeremešić, S., Pezo, L., Lončar, B., Krstić, Đ., ... & Ćupina, B. (2022). The effects of winter cover crops on maize yield and crop performance in semiarid conditions—Artificial neural network approach. *Agronomy*, *12*(11), 2670.
[20] Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. *Plos one*, *16*(6), e0252402.
[21] Fan, J., Bai, J., Li, Z., Ortiz-Bobea, A., & Gomes, C. P. (2022, June). A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 11, pp. 11873-11881).
[22] Agarwal, S., & Tarar, S. (2021). A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. In *Journal of Physics: Conference Series* (Vol. 1714, No. 1, p. 012012). IOP Publishing.
[23] Mkuhlani, S., Kephe, P. N., Rusere, F., & Ayisi, K. (2024). Modelling approaches for climate variability and change mitigation and adaptation in resource-constrained farming systems. *Frontiers in Sustainable Food Systems*, *8*, 1510162.
[24] Tantalaki, N., Souravlas, S., & Roumeliotis, M. (2019). Data-driven decision making in precision agriculture: The rise of big data in agricultural systems. *Journal of agricultural & food information*, *20*(4), 344-380.
[25] Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2020). Machine learning applications for precision agriculture: A comprehensive review. *IEE Access*, *9*, 4843-4873.
[26] Dehghanisanij, H., Emami, H., Emami, S., & Rezaverdinejad, V. (2022). A hybrid machine learning approach for estimating the water-use efficiency and yield in agriculture. *Scientific Reports*, *12*(1), 6728.
[27] Rezapour, S., Jooyandeh, E., Ramezanzade, M., Mostafaeipour, A., Jahangiri, M., Issakhov, A., ... & Techato, K. (2021). Forecasting rainfed agricultural production in arid and semi-arid lands using learning machine methods: A case study. *Sustainability*, *13*(9), 4607.
[28] luoPandit, P., Bakshi, B., & Gangadhar, V. (2021). Hybrid Time Series Models for Forecasting Maize Production in India. *Current Journal of Applied Science and Technology*, *40*(23), 49-57.
[29] Manjunath, M. C., & Palayyan, B. P. (2023). An efficient crop yield prediction framework using a hybrid machine learning model. *Revue d'Intelligence Artificielle*, *37*(4), 1057.
[30] Yan, Y., Wang, Y., Li, J., Zhang, J., & Mo, X. (2025). Crop Yield Time-Series Data Prediction Based on Multiple Hybrid Machine Learning Models.

Research Article

[31] Abdel-salam, M., Kumar, N., & Mahajan, S. (2024). A proposed framework for crop yield prediction using a hybrid feature selection approach and optimized machine learning. *Neural Computing and Applications*, *36*(33), 20723-20750.
[32] Nosratabadi, S., Imre, F., Szell, K., Ardabili, S., Beszedes, B., & Mosavi, A. (2020). Hybrid machine learning models for crop yield prediction. *arXiv preprint arXiv:2005.04155*.
[33] Hu, G., Archontoulis, S. V., & Huber, I. L. Coupling Machine Learning and Crop Modeling Improves Crop Yield Prediction in the US Corn Belt. In *ASA, CSSA, and SSSA International Annual Meetings (2020)| VIRTUAL*. ASA-CSSA-SSSA.
[34] Sarkar, S., Leyton, J. M. O., Noa-Yarasca, E., Adhikari, K., Hajda, C. B., & Smith, D. R. (2025). Integrating Remote Sensing and Soil Features for Enhanced Machine Learning-Based Corn Yield Prediction in the Southern US. *Sensors (Basel, Switzerland)*, *25*(2), 543.
[35] Iniyan, S., & Jebakumar, R. (2022). Mutual information feature selection (MIFS) based crop yield prediction on corn and soybean crops using multilayer stacked ensemble regression (MSER). *Wireless Personal Communications*, *126*(3), 1935-1964.
[36] Chen, Y., Zhang, Y., & Tan, Y. (2022). A comparative study of the cost–benefit strategy with the learning ensembles of decision stumps in polymetallic prospectivity modelling. *Earth Science Informatics*, 1-16.
[37] Ngo, G., Beard, R., & Chandra, R. (2022). Evolutionary bagging for ensemble learning. *Neurocomputing*, *510*, 1-14.
[38] Lolić, I., Sorić, P., & Logarušić, M. (2022). Economic policy uncertainty index meets ensemble learning. *Computational Economics*, *60*(2), 401-437.
[39] Le, H., Peng, B., Uy, J., Carrillo, D., Zhang, Y., Aevermann, B. D., & Scheuermann, R. H. (2022). Machine learning for cell type classification from single-nucleus RNA sequencing data. *Plos one*, *17*(9), e0275070.
[40] Zhang, L., Zhang, Z., Tao, F., Luo, Y., Cao, J., Li, Z., ... & Li, S. (2021). Planning maize hybrids adaptation to future climate change by integrating crop modelling with machine learning. *Environmental Research Letters*, *16*(12), 124043.
[41] Harsányi, E., Bashir, B., Arshad, S., Ocwa, A., Vad, A., Alsalman, A., ... & Mohammed, S. (2023). Data mining and machine learning algorithms for optimizing maize yield forecasting in central Europe. *Agronomy*, *13*(5), 1297.
[42] Luo, Y., Wang, H., Cao, J., Li, J., Tian, Q., Leng, G., & Niyogi, D. (2024). Evaluation of machine learning-dynamical hybrid method incorporating remote sensing data for in-season maize yield prediction under drought. *Precision Agriculture*, *25*(4), 1982-2006.
[43] Saravanan, K. S., & Bhagavathiappan, V. (2024). Prediction of crop yield in India using machine learning and hybrid deep learning models. *Acta Geophysica*, *72*(6), 4613-4632.
[44] Kenduiywo, B. K., & Miller, S. (2024). Seasonal Maize yield forecasting in South and East African Countries using hybrid Earth observation models. *Heliyon*, *10*(13).
[45] Haque, M. A., Marwaha, S., Deb, C. K., Nigam, S., Arora, A., Hooda, K. S., ... & Agrawal, R. C. (2022). Deep learning-based approach for the identification of diseases of the maize crop. *Scientific reports*, *12*(1), 6334.
[46] Abedinpour, M., Sarangi, A., Rajput, T. B. S., Singh, M., Pathak, H., & Ahmad, T. (2012). Performance evaluation of the AquaCrop model for the maize crop in a semi-arid environment. *Agricultural Water Management*, *110*, 55-66.
[47] Patil, Y., Ramachandran, H., Sundararajan, S., & Srideviponmalar, P. (2025). Comparative Analysis of Machine Learning Models for Crop Yield Prediction Across Multiple Crop Types. *SN Computer Science*, *6*(1), 64.
[48] Wang, Y., Huang, H., & State, R. (2024). Cross-Domain Early Crop Mapping using CropSTGAN. *IEEE Access*.
[49] Li, J., Zhang, D., Yang, F., Zhang, Q., Pan, S., Zhao, X., ... & Zhao, C. (2024). TrG2P: A transfer-learning-based tool integrating multi-trait data for accurate prediction of crop yield. *Plant Communications*, *5*(7).
[50] Habibi, L. N., Matsui, T., & Tanaka, T. S. (2024). Critical evaluation of the effects of a cross-validation strategy and machine learning optimization on the prediction accuracy and transferability of a soybean yield prediction model using UAV-based remote sensing. *Journal of Agriculture and Food Research*, *16*, 101096.
[51] Priyatikanto, R., Lu, Y., Dash, J., & Sheffield, J. (2023). Improving generalisability and transferability of machine-learning-based maize yield prediction model through domain adaptation. *Agricultural and Forest Meteorology*, *341*, 109652.

Research Article

[52] Nyborg, J., Pelletier, C., Lefèvre, S., & Assent, I. (2022). TimeMatch: Unsupervised cross-region adaptation by temporal shift estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, *188*, 301-313.
[53] Celisse, A. (2014). Optimal cross-validation in density estimation with the L^2-loss.
[54] Cravero, A., Pardo, S., Sepúlveda, S., & Muñoz, L. (2022). Challenges to using machine learning in agricultural big data: a systematic literature review. *Agronomy*, *12*(3), 748.
[55] Wadhwa, D., & Malik, K. Deep Learning Generalized Hybrid Models for Multi-Species Crop Disease Classification with Explainable Insights. *Available at SSRN 5139152*.
[56] Cartolano, A., Cuzzocrea, A., & Pilato, G. (2024). Analyzing and assessing explainable AI models for smart agriculture environments. *Multimedia Tools and Applications*, *83*(12), 37225-37246.
[57] Turgut, Ö., Kök, İ., & Özdemir, S. (2024, December). AgroXAI: Explainable AI-Driven Crop Recommendation System for Agriculture 4.0. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 7208-7217). IEEE.
[58] Khaki, S., Pham, H., & Wang, L. (2021). Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. *Scientific Reports*, *11*(1), 11132.
[59] Cheng, M., Penuelas, J., McCabe, M. F., Atzberger, C., Jiao, X., Wu, W., & Jin, X. (2022). Combining multiple indicators with machine-learning algorithms for maize yield early prediction at the county-level in China. *Agricultural and Forest Meteorology*, *323*, 109057.
[60] Das, P., Naganna, S. R., Deka, P. C., & Pushparaj, J. (2020). Hybrid wavelet packet machine learning approaches for drought modeling. *Environmental Earth Sciences*, *79*(10), 221.
[61] QianChuan, L. I., ShiWei, X. U., ZHANG, Y., ZHUANG, J., DengHua, L. I., BaoHua, L. I. U., ... & Hao, L. I. U. (2024). Stacking ensemble learning, modeling, and forecasting of maize yield based on meteorological factors. *Scientia Agricultura Sinica*, *57*(4), 679-697.
[62] Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., Borodulin, A., & Tynchenko, Y. (2024). Predicting sustainable crop yields: Deep learning and explainable AI tools. *Sustainability*, *16*(21), 9437.
[63] Khlif, M., Chahbi Bellakanji, A., Escorihuela, M. J., Sánchez Alcalde, G., & Lili Chabaane, Z. Automatic Early Multi-Year Cereal Yield Prediction in Semi-Arid Regions: Integrating Machine Learning and Satellite Drought Indices. *Maria José and Sánchez Alcalde, Guillem and Lili Chabaane, Zohra, Automatic Early Multi-Year Cereal Yield Prediction in Semi-Arid Regions: Integrating Machine Learning and Satellite Drought Indices*.
[64] Jia, W., Wei, Z., & Zhang, L. (2022). A novel prediction and planning model for the benefit of irrigation water allocation based on deep learning and uncertain programming. *Water*, *14*(5), 689.