A Review and Analysis of Machine Learning Algorithms for Sustainable and Data Driven Crop Yield Prediction in Modern Agriculture

Sushmitha N[1*], Dr. Kendaganna Swamy S[2]

[1]Assistant Professor, Department of Information Science and Engineering, RV College of Engineering,*
[2]Assistant Professor, Department of Electronics and Instrumentation Engineering, RV College of Engineering.

Abstract: *Agriculture is the backbone of India's economy. It sustains over half of the population and plays a vital role in the country's Gross Domestic Product (GDP). The agriculture sector faces significant uncertainty in balancing the demand and supply, largely due to the limited use of techno logy. The challenges like unpredictable weather patterns, soil degradation and limited resources makes crop yield prediction difficult for farmers as they often lack the necessary knowledge. To overcome this difficulty and also to ensure enough productivity in agriculture, it is important to accurately predict the crop yields. Traditional methods, such as statistical and agricultural models, have been helpful, but they often do not provide the high level of accuracy needed for modern farming. Recent advancements in machine learning (ML) have demonstrated significant potential in enhancing the accuracy of crop yield prediction models. The various Machine learning techniques offer an innovative solution, enhancing prediction accuracy through advanced data analysis that synthesizes historical yields, weather patterns and soil characteristics. This paper reviews contemn porary ML techniques for crop prediction, providing insights and recommendations for applying to advance the sustainability and productivity of Indian agriculture.*

Keywords: *Agriculture, Crop, Yield Prediction, Crop Prediction, Machine Learning, Data Analytics, Soil Characteristics.*

1. Introduction

India's extensive geographical diversity plays a crucial role in sustaining its agricultural sector, which remains a fundamental component of the national economy. As of 2025, agriculture supports the livelihoods of nearly 43.5% of the workforce and contributes approximately 16% to the Gross Domestic Product (GDP) [1]. Crops such as rice, wheat, sugarcane, turmeric, and onions are extensively cultivated across the country's varied climatic regions, serving as critical sources of rural employment and national food security [2]. However, agricultural growth has recently decelerated, with a growth rate of just 1.4% during the 2023–2024 financial year, primarily attributed to climate instability and irregular monsoons [3]. According to final estimates from the Ministry of Agriculture &Farmers Welfare, India's total foodgrain production reached a record 332.22 million tonnes in the 2023-24 crop year, surpassing the previous year's output of 329.6 million tonnes. Notably, rice and wheat production also hit record levels at 137.82 million tonnes and 113.29 million tonnes, respectively. Despite these achievements, persistent challenges such as unpredictable weather patterns, water scarcity, and soil degradation underscore the urgent need for more reliable and data-driven crop prediction methods [4]. The percentage of India's population dependent on agriculture, as shown in Figure 1, has steadily declined over the years [5], reflecting a broader trend of economic diversification. With increasing urbanization and a shift towards other industries, this decline has contributed to the growth of India's GDP [6]. However, as the population continues to grow, it becomes crucial to implement sustainable agricultural strategies. Farmers are increasingly losing interest in agriculture due to low and unstable incomes, rising input costs, and heavy debt burdens. Climate change, unpredictable weather patterns, and water scarcity have made farming riskier. Furthermore, the lack of access to modern technology, inadequate infrastructure, and fragmented landholdings significantly reduce productivity. As a result, many farmers migrate to urban areas in search of better employment opportunities and stable livelihoods [6]. Hence, focused efforts are essential not only to support the economic well-being of farmers but also to ensure that agriculture remains a cornerstone of the nation's economy and food security.

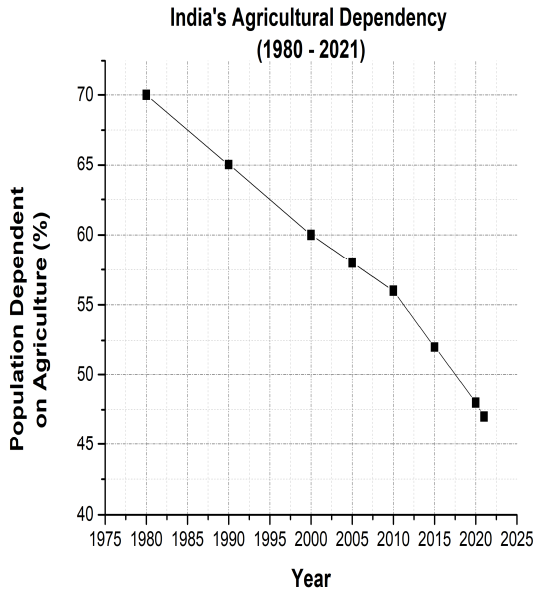

Figure 1: Decline in India's Agricultural Dependency (1980-2021).

In this context, the traditional methods of forecasting crop yield have relied heavily on statistical techniques. These methods analyse past yield records, applying models like time series and regression analysis to predict future productivity [7]. Farmers have also used soil testing, seasonal trends, and meteorological data to improve crop health and resource allocation. However, while these methods are useful, they often lack the precision required to respond to sudden shifts in environmental conditions, underscoring the need for advanced approaches [8]. The advanced Machine learning has emerged as a key technology in improving the accuracy of crop predictions. By analysing large data sets including historical yield records, soil quality and weather patterns, ML identifies the trends and forecasts the outcomes. Algorithms like linear regression, decision trees and neural networks reveal important links between environmental factors such as rainfall and yield levels. These models continuously adapt in real time, integrating new data to refine predictions and respond to changing weather, pest outbreaks, and market fluctuations [9]. Beyond yield prediction, ML applications in agriculture enhance resource efficiency, risk management, and sustainable practices. Using real-time sensor-based data from IoT devices, ML-based decision systems help farmers to determine optimal planting times and manage water and fertilizer usage effectively. These systems also enable early detection of crop stress or disease, reducing losses. By simulating various weather patterns, ML models help farmers prepare for unfavourable conditions, building resilience across agricultural practices [10].Machine learning (ML) is also playing a key role in promoting sustainable farming by enabling precise adjustments based on water quality, soil health, and climate data [11]. These targeted interventions help minimize the overuse of fertilizers and pesticides, thereby supporting both crop productivity and environmental health. As India continues to face critical challenges such as water scarcity and soil degradation, adopting a data-driven approach is essential for establishing sustainable agricultural practices that preserve natural resources for the future [12].By integrating ML tools into everyday farming operations, Indian agriculture is becoming more adaptive, efficient, and environmentally conscious. As illustrated in Figure 2, machine learning algorithms are typically categorized into three major types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning relies on labelled datasets, where each input is matched with a known output. Through this process, the algorithm learns to map inputs to outputs, enabling it to make accurate predictions on new, unseen data. Within supervised learning, classification algorithms are used to predict categorical outcomes, while regression algorithms are applied to forecast continuous numerical values [13].

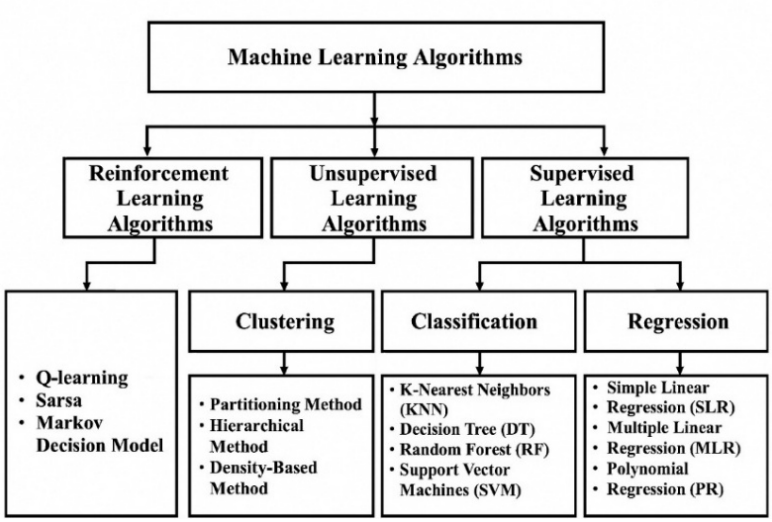

Figure 2: Classification of Machine Learning Algorithms and Their Functional Categories.

In contrast, unsupervised learning works with unlabelled data to identify patterns or group similar data points. These algorithms aim to uncover hidden structures within the data without predefined labels, making them more complex to implement than supervised methods. Reinforcement learning, on the other hand, involves algorithms learning through interaction with their environment. By receiving rewards or penalties based on their actions, the algorithm iteratively improves its decision making over time [14]. Looking ahead, technology's role in agriculture is expected to expand further through initiatives focused on strengthening digital infrastructure, increasing government support, and educating farmers on data-driven practices [15]. Programs such as digital rural outreach, agritech policies, and collaborations with private organizations are enhancing farmers' access to ML tools, empowering even small-scale farmers to boost productivity and reduce losses. With broader adoption, these advancements are set to contribute significantly to India's self-sufficiency in food production, while fostering a more innovative, resilient, and sustainable agricultural sector. As machine learning and related technologies continue to evolve, Indian agriculture is poised to become a global model of innovation [16]. The upcoming sections discuss the core machine learning algorithms along with their mathematical foundations (Section 2), provide a comprehensive review of related literature (Section 3), and conclude the study (Section 4). This research work centres on advancing crop yield prediction techniques to boost farm productivity and profitability globally. Additionally, it underscores how predictive models can guide sustainable farming practices and support data-driven decisions in agriculture.

2. Machine Learning Algorithms

This section explores essential machine learning algorithms, focusing on their underlying mathematical concepts, how they learn from data, and their application in analysing agricultural datasets. These techniques are crucial for making accurate predictions and supporting smart, data driven decisions in agriculture.

2.1 Linear Regression

Linear regression is a supervised learning method used to predict continuous outcomes by establishing a relationship between one or more independent variables and a dependent variable. The model is represented mathematically as in equation (1).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \quad (1)$$

where Y is the dependent variable, $X_1, X_2, ..., X_n$ are the independent variables, β_0 represents the intercept, and $\beta_1, \beta_2, ..., \beta_n$ are the regression coefficients. The term ε denotes the error term, accounting for variations not explained by the model. To estimate the optimal values of these coefficients, the *Ordinary Least Squares (OLS)* method is employed. This technique minimizes the sum of squared differences between actual and predicted values. The coefficients are derived using equation (2).

$$\beta = (X^T X)^{-1} X^T Y \quad (2)$$

where X^T is the transposed matrix of the independent variables, and $(X^T X)^{-1}$ represents the inverse of the product of X^T and X. This calculation ensures that the regression model determines the best-fitting line, thereby reducing prediction errors and improving accuracy.

2.2 Logistic Regression

Logistic regression is a supervised learning method designed for classification tasks where the target variable is categorical rather than continuous. It estimates the probability that a given input belongs to a specific class by applying the sigmoid function to a linear combination of input variables. The logistic regression model is mathematically expressed as in equation (3).

$$P(Y = 1 \mid X) = 1 / (1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}) \quad (3)$$

where $P(Y = 1 \mid X)$ represents the probability of the dependent variable Y belonging to class 1, and $X_1, X_2, ..., X_n$ denote the independent variables. The parameters $\beta_0, \beta_1, ..., \beta_n$ are estimated using Maximum Likelihood Estimation (MLE), which identifies the values that maximize the likelihood of the observed data. The optimization of these coefficients is carried out through Gradient Descent, which updates them iteratively using equation (4).

$$\beta_j = \beta_j - \alpha \sum (y - \hat{y}) X_j \quad (4)$$

where α is the learning rate, and $\hat{y}$ represents the predicted probability. This iterative process refines the model parameters, enabling the model to effectively distinguish between classes by mapping input features to corresponding probabilities.

2.3 Decision Tree

A *decision tree* is a supervised learning technique that predicts outcomes by recursively partitioning the dataset based on feature values. It is composed of nodes, which represent decision points, branches that indicate possible outcomes, and leaf nodes that provide the final classification or prediction. At each step, the algorithm determines the most informative feature for splitting the data using impurity measures such as Entropy (Information Gain) or the Gini Index. The Entropy measure, which quantifies the uncertainty or impurity in a dataset, is calculated using equation (5)

$$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (5)$$

where p_i represents the fraction of samples belonging to class i. Information Gain is then computed as the reduction in entropy after the dataset is split based on a particular feature. Another widely used metric for evaluating data purity is the Gini Index, which is given in equation (6).

$$\text{Gini} = 1 - \sum_{i=1}^{c} p_i^2 \quad (6)$$

A lower Gini Index indicates a better split, meaning the data is more homogeneous after partitioning. The decision tree algorithm iteratively applies these calculations to construct a structured tree that efficiently classifies new data points.

2.4 Random Forest

Random Forest is a supervised learning algorithm that enhances the reliability and accuracy of decision trees by generating multiple trees and combining their outputs. Each tree in the forest is trained on a randomly selected subset of the dataset, and the final prediction is determined using majority voting for classification tasks or averaging in the case of regression. The final prediction is mathematically expressed as in equation (7).

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} T_i\,(X) \tag{7}$$

where $T_i(X)$ denotes the prediction made by the *i*-th decision tree, and N represents the total number of trees in the model. By aggregating the results from multiple trees, Random Forests minimize overfitting and enhance overall predictive accuracy.

2.5 K-Means

K-Means is an unsupervised learning algorithm that groups data into K clusters by iteratively assigning each data point to the closest centroid and updating the centroids accordingly. The algorithm determines cluster assignments by calculating the Euclidean distance between each data point and the cluster centre, as shown in equation (8).

$$d = \sum (x_i - c_j)^2 \tag{8}$$

where *d* represents the distance, x_i is a data point, and c_j denotes the centroid of cluster j. After assigning the data points to clusters, the centroids are updated using equation (9).

$$C_j = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{9}$$

where *N* is the total number of points in the cluster. This process iterates until the centroids stabilize, ensuring that the clusters are well-defined and intra-cluster variance is minimized.

2.6 Hierarchical Clustering

Hierarchical clustering is an unsupervised learning technique that organizes data into a hierarchical structure of nested clusters. It follows either a **bottom-up (*agglomerative*)** approach, where individual data points are progressively merged into larger clusters, or **a top-down (*divisive*)** approach, where a single cluster is repeatedly divided into smaller groups. The similarity between clusters is measured using Euclidean Distance, calculated as shown in equation (10).

$$d(A, B) = \sum (A_i - B_i)^2 \tag{10}$$

where *A* and *B* represent the two clusters being compared. The clustering process continues iteratively by either merging similar clusters or splitting larger ones until the desired hierarchical structure is achieved. These fundamental categories of machine learning serve as the basis for a wide array of algorithms utilized across diverse application domains. In the agricultural sector, they play a pivotal role in facilitating the analysis of large and complex datasets, enabling data-driven decision-making, uncovering actionable insights, and ultimately enhancing productivity and efficiency. The selection of an appropriate algorithm is contingent upon the characteristics of the data and the specific objectives of the prediction or classification task.

The above section discussed various machine learning algorithms available for making predictions based on data. These algorithms play a crucial role in processing large datasets, identifying patterns, and generating accurate forecasts. They are widely used across different domains to enhance decision-making and optimize performance. The following section provides a concise over view of various research studies that have explored the application of machine learning algorithms in crop yield prediction worldwide. The focus is on how these algorithms have been utilized by researchers to develop machine learning models aimed at improving agricultural productivity and profitability.

3. Background study

In this research, a comprehensive study was conducted to survey of an agri-advisory system aimed at predicting crop yields for crops such as sugarcane, turmeric, onion, rice, paddy, and various other crops. Several research articles were reviewed and performed a thorough literature survey on crop prediction using machine learning techniques. Although many research articles address crop prediction, we identified only a select few that align with our problem statement, which are outlined below as part of the overall system development. Researchers, including Jharna Majumdar et al. (2017) [17], conducted a study examining crop yields and overall rainfall across India from 2003 to 2014. Their findings indicate that the meth ods employed in their analysis were insufficient to effectively address the nonlinear relationships among the various factors considered. This limitation suggests a need for more advanced analytical techniques that can better capture the complexities inherent in agricultural data, particularly in understanding how different variables interact to influence crop yields. Research utilizing a multivariable analytical approach was conducted by Ruchita Thombare et al. (2017) [18] the authors selected various factors, categorizing them into explanatory and response variables. However, it is evident that the methodology employed in their study was not validated, raising concerns about the reliability of the findings. This lack of validation highlights the import ance of ensuring methodological rigor in research to substantiate the conclusions drawn from such analyses. Hossein Aghighi et al. (2018) [19] investigate advanced machine learning models, including, Support Vector Regression (SVR), Random Forest Regression (RFR), Gaussian Process Regression (GPR) and Boosted Regression Tree (BRT), to predict silage maize yield using NDVI time-series data from Landsat 8. The study focuses on addressing the challenges of complex and inconsistent data in crop yield prediction. By averaging NDVI values across fields and performing 100 iterations per model, the research assesses both predictive accuracy and model stability. BRT demonstrated the highest accuracy, consistently achieving an R value above 0.87, while RFR showed notable stability, particularly in predicting yields across different years with minimal variance. These machine learning techniques outperformed traditional regression methods by effectively handling high-dimensional, variable data, highlighting the promise of ML in agricultural forecasting. The authors suggest further application of these methods to different crop types and regions to explore their broader relevance and potential in supporting agricultural decision-making. A. S. Terliksiz et al., (2019) [20] introduced hybrid deep learning (DL) models for crop prediction to highlight the benefits of using hybrid approaches. The models incorporate Deep Neural Networks (DNN), Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for predicting crop yields, demonstrating favourable outcomes compared to traditional machine learning algorithm. Accurate agricultural price prediction is crucial for maintaining the global economy. According to author, given the limited availability of time series price data, deep learning models, particularly LSTM, are employed for one-month price predictions in the agricul tural commodities sector. R.M.S.M.Rathnayake et.al (2019) [21] study presents the development of the Organic Cultivation Management and Prediction System (OCMPS), a software platform designed to support sus trainable organic farming in Sri Lanka. It assists organic farmers by predicting harvests, pricing, and optimal crop selection, integrating machine learning and optimization techniques. Additionally, OCMPS uses block chain for authenticating organic producers, ensuring consumers receive verified organic products. Serving as both a decision-support tool and an electronic marketplace for organic foods, this platform enhances farmer knowledge and fosters trust through direct partnerships among stakeholders, addressing a current gap in organic agriculture support systems in Sri Lanka.

Ranjini B Guruprasad et al. (2019) [22] examine paddy crop yield estimation using weather and soil data, with applications in optimizing supply chains, adjusting farming practices, price fore casting, and agricultural insurance risk assessment. Models based on satellite and weather data offer scalability across various crops and regions without needing specific crop or farming details. This study focuses on paddy yield estimation at district (coarser spatial resolution) and taluk (finer spatial resolution) levels in India, analysing model accuracy with different feature sets and machine learning techniques. It also presents a dis-aggregation method to use district-level yield data for predicting yields at the taluk level. Results show that taluk-level predictions from district data yielded an average error of 6% and a maximum error of 25%, with the best accuracy achieved using taluk data averaged weekly, which had an average error of 3.14% and a maximum error of 9.66%. Alaa Adel Araby et al. (2019) [23] investigate the use of Internet of Things (IoT) technology in precision agriculture by introducing a smart system that integrates IoT and machine learning to predict late blight disease in potatoes and tomatoes. This system utilizes a network of sensors to collect key data, such as temperature, relative humidity, leaf wetness, and soil moisture, providing farmers with timely alerts for pesticide application. This approach supports cost reduction and minimizes environmental impact. Data is sent via Wi-Fi and MQTT protocol to a gateway, where machine learning algorithms analyse the information to provide notifications displayed on an accessible website. Future enhancements may enable the system to achieve full automation, allowing the gateway to autonomously perform actions such as field spraying, thus improving the efficiency of crop management. Sukanya Randhawa et al. (2019) pro [24] poses a machine learning-based method for global crop identification that utilizes limited microwave and radar data, concentrating on the Corn Belt in the United States. The research focuses on recognizing unique crop signatures that remain consistent across various geographical locations, leading to the development of a dependable and scalable crop identification model specifically for corn and soybeans. The pertained model achieved notable accuracy levels, reaching 93% within the corn and soybean belt and 84% at its periphery, all at a spatial resolution of 20 meters. The results indicate that implementing effective strategies to address regional variability can enable the development of a universal crop identification model, and while the current approach effectively identifies corn and soybeans, it can also be readily adapted for other crop types.Y. Jeevan Nagendra Kumar et al. (2019) [25] explores the critical role of machine learning (ML) in predicting crop yields, utilizing supervised learning to establish connections between various input factors such as pH, temperature, rainfall, humidity and crop type and their corresponding yields. The research employs the Random Forest algorithm, which effectively forecasts crop yields based on historical data, providing essential insights into the best appropriate crops for distinct environmental circumstances. The study highlights the algorithm's capability to generate reliable results while maintaining ease of use, making it particularly beneficial for the agricultural sector. This research introduces a predictive system that integrates historical data and data mining methods, employing the Random Forest algorithm to enhance the accuracy of crop yield predictions. Increased precision in these forecasts translates into improved profitability for farmers by equipping them with crucial information regarding market demands and crop pricing. As a result, the system empowers farmers to make informed decisions about which crops to cultivate, optimizing yield and harvesting efficiency across a wide array of crop types. Nikhil R et al. (2020) [26] present a smart agriculture system that integrates IoT and machine learning (ML) to enhance irrigation, crop selection, and farm security, thereby addressing the inefficiencies of traditional farming practices. The system automates water sprinkler operations by monitoring soil moisture levels, ensuring that only the necessary amount of water is supplied, which promotes conservation. It utilizes ML algorithms to predict suitable crops based on soil conditions, aiding farmers in making informed crop selection decisions. Additionally, IoT-enabled sensors and ultrasonic alarms help deter wild animal intrusions, improving farm security. The system also sends alerts via SMS and email to keep farmers updated on field conditions. This cost-effective solution enables real-time decision-making, boosts irrigation efficiency, and increases crop yield while maintaining an environmentally friendly approach. Rosemary Nalwanga et al. (2021) [27] introduced an embedded machine learning (ML) system designed for precision agriculture, emphasizing environmental preservation. In contrast to conventional methods that depend heavily on fertilizers, this system forecasts the best crops to grow by analysing real-time soil

and weather data, thereby reducing fertilizer reliance and promoting soil health. It gathers daily soil metrics, integrates weather forecasts, and utilizes embedded ML to suggest crops that match the current soil conditions, allowing for immediate monitoring of soil nutrients. This embedded system effectively addresses connectivity issues, particularly in areas with poor internet access, by processing information locally instead of relying on cloud services. The findings demonstrated comparable accuracy in both cloud-based and embedded models, highlighting the potential of embedded AI in precision agriculture. The research also confirmed that synthetic data can be beneficial when there is a scarcity of training data. Future efforts will aim at fully implementing this system, which is anticipated to boost crop production while minimizing environmental damage and fertilizer expenses, thereby fostering sustainable agricultural. Munavvara Tahaseen et al. (2021) [28] review the implementation of machine learning (ML) algorithms and advanced sensor technologies in agriculture, emphasizing their roles in early detection of plant diseases, crop yield forecasting, and managing pests and weeds. This research discusses various key ML techniques, such as Support Vector Machines (SVMs), Bayesian classifiers, Ada Boost, Random Forest, and several Convolutional Neural Network (CNN) architectures, including AlexNet and GoogLeNet, particularly for tasks like leaf analysis and disease identification. While the study acknowledges the effectiveness of these algorithms, it also highlights that much of the existing research relies on controlled datasets rather than real-world scenarios. Furthermore, the research work examines the significance of remote sensing and hyperspectral data in precision agriculture for disease detection and yield prediction, emphasizing the transformative potential of these technologies in modern agricultural practices. Anandhan K. et al. (2021) [29] explore the transformative capabilities of advanced machine learning (ML) and deep learning (DL) technologies in aiding Indian farmers to maximize crop yields, particularly in rice production. Traditionally, Indian farmers depend on experiential methods for crop planning, but the growing availability of digital data presents challenges in effectively managing and interpreting this information. This research article investigates various ML and DL techniques for classifying and segmenting paddy leaf diseases, with the goal of improving yield through early disease detection and actionable insights. Additionally, the study addresses the challenges related to data scarcity affecting many existing models, especially those pertaining to paddy and wheat, which are staple crops in India. The authors of this research conclude with a call for future research aimed at developing comprehensive datasets from local farms to enhance predictive models, ultimately promoting productivity through smart agricultural practices.Francis Muthoni et al. (2021) [30] examine the application of machine learning and remote sens ing data to evaluate the effectiveness of conservation agriculture (CA) compared to conventional practices (CP) in smallholder maize farming across four southern African countries over a 13-year period. Utilizing a random forest algorithm to analyze spatial and temporal yield data, the study reveals that CA can provide significant benefits, especially during drought years, with yield increases of up to 1 t/ha compared to CP. However, during years of above-normal rainfall, CA led to minor yield losses, with Mozambique being an exception. The analysis identified key predictors of yield, including precipitation, temperature, altitude, and soil conditions, emphasizing the critical role of rainfall timing in essential maize growth phases. The resulting maps offer guidance for spatially targeted CA practices by identifying regions where CA improves drought resilience, thus serving as an efficient and cost-effective tool for sustainable agricultural planning in varied socio-economic and biophysical contexts. Dushyant Kumar Singh et al. (2021) [31] investigate the role of Precision Agriculture, enhanced by technologies such as IoT, Wireless Sensor Networks (WSN), and Machine Learning (ML), in modernizing farming practices to optimize resource utilization, particularly in irrigation. Often referred to as Agriculture 4.0 when integrated with IoT, Precision Agriculture facilitates real-time monitoring and control of agricultural parameters, including soil moisture, crop health, and irrigation requirements. ML contributes to predicting and adapting to the needs of farmland, thus improving irrigation efficiency a critical objective given that agriculture accounts for nearly 70usage. This study reviews the effectiveness of various ML algorithms and WSN communication protocols, highlighting LoRa as a power-efficient option for WSN nodes. However, the implementation of these technologies faces several challenges, such as security concerns, gaps in technical knowledge, environmental resilience, reliability, and issues related to connectivity and scalability. These challenges underscore the necessity for robust solutions to enhance the effectiveness

of IoT and WSN in Precision Agriculture. G. Mariammal et al. (2021) [32] focus on predicting crop cultivation by introducing a novel feature selection (FS) method known as Modified Recursive Feature Elimination (MRFE). The proposed method of this research work identifies key soil and environmental factors, such as rainfall, temperature, and fertilizer application that affect crop suitability in different regions. By ranking significant features, MRFE enhances prediction accuracy, aiding farmers in choosing the most appropriate crops for their specific locations. The MRFE was evaluated using several machine learning classifiers, including kNN, Naive Bayes, Decision Trees, SVM, Random Forest, and a bagging classifier, achieving an accuracy of 95%, which surpasses that of other FS techniques. The authors of this research article states that, the MRFE combined with the bagging classifier was particularly effective in predicting suitable crops across various datasets, although further optimization is necessary for handling larger datasets. Akshay Kumar Gajula et al. (2021) [33] emphasizes the importance of agriculture in driving national economic growth while highlighting the challenges to crop yields from climate variability, outdated farming techniques, and insufficient irrigation systems. The study employs Machine Learn ing (ML) techniques, specifically the K-Nearest Neighbors (KNN) algorithm, to evaluate soil quality and predict suitable crops for cultivation based on temperature and soil conditions. Results indicate that this method is effective in guiding crop selection and yield predictions, offering farmers valuable insights for optimizing their practices and boosting production. Although the model achieves high accuracy, it is limited by a small dataset and does not consider potential climatic disasters. Future enhancements may include the integration of geospatial analysis to provide richer data and improve the accuracy of predictions. Sharma et al. (2022) [34] underscore the significance of agriculture as the backbone of the In dian economy and stress the necessity for crop prediction techniques to meet farmers' expected demands. The authors of this research article identify key challenges in recommending crops and predicting yields, leading to the development of two models for crop prediction. One part of the model focuses on forecasting crop yields by analyzing factors such as geo-climatic conditions, soil type, district, season, and crop variety. This approach is intended to support both the government and farmers in making strategic decisions on agricultural risk management and pricing, aiming to maximize profits. For crop suggestions, the model utilizes various machine learning algorithms, including the Decision Tree Classifier, Naive Bayes Classifier, KNN Classifier, Random Forest Classifier, Gradient Boosting, and XGBoost. The study ultimately recommends the Random Forest Regressor for predicting crop yields, achieving an accuracy of 89%, and the Random Forest Classi f ier for crop suggestions, with an accuracy of 98%. Thirumal et al. (2023) [35] emphasize the critical role of crop yield prediction in agriculture, particularly for paddy and rice crops. The study notes that while several methods for predicting crop yields exist, they often fail to deliver accurate results for paddy, making it a challenging task for farmers. The authors identify key factors contributing to the uncertainty in crop yields, such as unpredictable weather conditions, varying land types, and resource availability. To address these challenges, the paper introduces an automated approach for rice crop yield prediction using the Sine Cosine Algorithm with a Weighted Regularized Extreme Learning Machine (SCA-WRELM). This method aims to improve rice productivity forecasting by normalizing data through a min-max process to ensure consistency. The SCA-WRELM model then uses the WRELM framework to provide more accurate predictions of rice crop yields.Deshmukh et al. (2023) [36] underscore the importance of agricultural prediction for bolster ing a nation's economy. Their research identifies key factors, including soil composition, water availability, temperature, and climate change, which significantly influence agricultural outcomes. Given the complexities involved, there is growing interest in using Machine Learning (ML) to assist farmers in achieving optimal crop yields. The article proposes the development of a crop selection model powered by ML, drawing on a comprehensive database of crop-specific data, local weather, soil quality, and water availability. This model acts as a decision-support system, enabling farmers to choose crops best suited to their region's environmental conditions while also incorporating methods to monitor crop growth for greater agricultural efficiency. By analyzing climate patterns, soil properties, and historical yields, the ML model can accurately predict crop yields and help farmers strategically plan cultivation across different areas. According to the authors, this approach holds great potential for improving resource management and productivity, offering a robust tool for both farmers and policymakers to optimize agricultural

practices. In their 2023 study, T. Sathies Kumar et al. [37] explore the transformative role of Machine Learning (ML) in supporting sustainable agriculture, especially in light of climate change and population growth. The research underscores the limitations of traditional farming methods and highlights how ML's predictive power enables more efficient, data-driven crop management to boost yields and reduce environmental impact. By employing algorithms like decision trees, neural networks, satellite imagery and leveraging real-time data from IoT sensors, the study highlights how machine learning can enhance irrigation, fertilization, and pest management practices for greater efficiency and effectiveness. It advocates for the ethical and widespread adoption of ML in agriculture to bolster food security globally and advance sustainable farming, creating a resilient, productive future for agricultural systems.Rajendra. Pawar et al. (2023) [38] propose a machine learning-based framework designed to aid farmers and agricultural researchers in identifying the best crops to cultivate by considering various environmental factors, including rainfall, soil type, temperature and humidity. The model employs K-means clustering to organize environmental data and utilizes classification algorithms such as support vector machines (SVMs) or decision trees for crop prediction, providing a data driven solution to enhance crop yields. The approach has been validated through historical yield data and practical experiments, showcasing its effectiveness. The study concludes that machine learning can significantly optimize crop selection, enabling farmers to make informed decisions that increase yields, minimize waste, and lessen dependency on chemicals, thereby fostering both economic sustainability and environmental health in agriculture. Sunil G L et al. (2024) [39] underscore the critical role of agriculture in India, driven by the increasing population and food demand. The authors advocate for leveraging advanced technologies, such as machine learning, data mining, and remote sensing, to enhance crop yield predictions. Their comparative analysis of various machine learning techniques reveals that supervised classification methods significantly outperform regression and unsupervised techniques in accurately forecasting yields across different crops. While classification methods utilize a wider array of parameters and achieve superior accuracy, the study also highlights the untapped potential of clustering methods, particularly bi-clustering in unsupervised learning, suggesting avenues for future research to tackle the complexities of crop yield prediction. Vanitha K et al. (2024) [40] present an innovative method for predicting crop yields in Indian agriculture by integrating power transformations, specifically the Yeo-Johnson transformation, into machine learning models. This technique effectively addresses non-linear distribution issues prevalent in agricultural datasets, leading to enhanced predictive accuracy. The study utilizes a comprehensive dataset encompassing variables such as climate, soil properties, crop types, and historical yields, and compares various machine learning models, including Linear Regression, Gradient Boosting and Random Forest. The results reveal that the Yeo-Johnson transformation improves performance across all models, with Gradient Boosting achieving the highest level of accuracy. This research underscores the importance of advanced preprocessing techniques like power transformations in improving model accuracy, which in turn aids better decision-making for farmers, policymakers, and researchers. The findings emphasize the transformative potential of data preprocessing in agricultural applications and suggest future possibilities for integrating real-time data to further advance agricultural forecasting and sustainability in India. Basavaraju N M et al. (2024) [41] introduce the Smart Agriculture Yield and Fertilizer Optimization System (SAYFOS), an innovative solution designed to enhance crop productivity and resource man agement through the incorporation of IoT, data analytics, and machine learning. SAYFOS facilitates real-time monitoring of soil conditions and crop health by analyzing data from soil sensors, weather forecasts, and satellite imagery, allowing for optimized fertilizer application and accurate crop yield predictions. The results indicate that SAYFOS improves crop yield by 0.25%, reduces water consumption by 0.33%, and enhances fertilizer efficiency by 0.25% when compared to traditional farming methods. By minimizing fertilizer overuse and improving decision-making for farmers, this system supports sustainable agricultural practices. SAYFOS highlights the transformative potential of incorporating advanced technologies in agriculture, paving the way for more sustainable, efficient, and productive farming practices. Table 1 provides an overview of various machine learning algorithms developed for crop prediction, highlighting the challenges and complexities in the field. Although numerous models exist, there remains a pressing need for new and improved approaches due to several

factors. Agriculture is inherently complex, influenced by diverse variables such as climate, soil properties, pest dynamics, and farming practices. This complexity demands models capable of capturing intricate and nonlinear relationships.

Table 1: Comparisons of various ML algorithms for crop prediction

Author Name	Year	Crops Used	Methods Used	Model Accuracy
Hossein Aghighietal.	2018	Wheat, Maize, Soybean,Sugar cane,Cotton	Random Forest, SVM	R^2 = 0.80-0.88, RMSE = 0.5-1.2
		Rice	ANN, Decision Trees	R^2 = 0.80-0.90, RMSE = 0.4-1.0
		Soybean	Gradient Boosting, LSTM	R^2 = 0.86-0.92, RMSE = 0.3-0.8
		Sugarcane	Linear Regression, XGBoost	R^2 = 0.78-0.85, RMSE = 1.0-2.5
		Cotton	KNN, Ensemble Methods	R^2 = 0.82-0.88, RMSE = 1.0-2.5
R.M.S.M. Rathnayake et al.	2019	Multiple organic crops	Linear Regression, Knapsack Optimization	Harvest: 94.01%, Price: 90.91%
Alaa Adel Araby	2019	Potatoes, Tomatoes	IoT with sensors, MQTT, Polynomial Regression, ML	99.97% (9th order polynomial regression)
Sukanya Randhawa et al.	2019	Corn, Soybeans	Random Forest, SAR Data	Binary classification: 92.6% (Corn), 96.63% (Soy)
Y. Jeevan Nagendra Kumar et al.	2019	Multiple crops	Random Forest	Better accuracy than Decision Tree and SVR
Nikhil R	2020	Ragi (example)	SVC, CNN	Not Mentioned
Rosemary Nalwanga et al.	2021	Maize, Beans, Lentil, Peas, Watermelon	Neural Network Classifier	92.2% (Synthetic Data), 78.8% (Non-Synthetic)
Francis Muthoni et al.	2021	Maize (18 varieties)	CIMMYT Agronomic Data, Remote Sensing, Random Forest	R^2 = 0.63, RMSE ±1.2 t/ha
Dushyant Kumar Singh et al.	2021	Maize, Rice, Wheat	Random Forest, Remote Sensing, SVM	R^2 = 0.63, RMSE ±1.2 t/ha
G. Mariammal et al.	2021	9 crop classes	Feature Selection, kNN, NB, DT, SVM, RF	MRFE + Bagging achieved highest accuracy
Ankita Sharma et al.	2022	Various Crops	Random Forest Regressor, Decision Tree Regressor	Model Accuracy = 89%, High R^2, Low RMSE
T. Sathies Kumar et al.	2023	Generalized Crop	Decision Trees, Random Forests, Neural Networks	Not Mentioned
Dr. Rajendra Pawar et al.	2023	Corn, Wheat, Soybeans, Cotton, Potatoes, Rice, Sugarcane	K-means Clustering, Logistic Regression, Decision Trees, SVM	Not Mentioned

Moreover, many existing models struggle to handle the vast and heterogeneous nature of agricultural data, including unstructured sources like satellite imagery and weather records. This underscores the need for models that can effectively integrate and process diverse datasets. Enhancing predictive accuracy is another key motivation, as recent advancements in deep learning and other sophisticated techniques offer the potential to surpass traditional methods. The impacts of climate change and evolving agricultural practices further necessitate adaptable models that can respond to emerging environmental patterns. Additionally, developing models with intuitive interfaces and interpretable outputs can empower farmers and policymakers with actionable insights. Tailoring these models to specific regional conditions can also lead to more localized and relevant predictions.

4. Conclusion

The agricultural sector is grappling with several challenges in maximizing crop yields, including unpredictable climate patterns, soil degradation, and limited resources. Despite the advancements in crop yield prediction algorithms, many of these tools are not easily accessible or user-friendly, hindering farmers from effectively utilizing advanced technologies to improve their practices. This research article seeks to fill gap by conducting an extensive analysis of existing studies on crop yield prediction methods. It reviews various machine learning and deep learning algorithms currently in use, highlighting emerging trends and innovations in the field. By incorporating insights from various studies, this research aims to offer a clearer understanding of the capabilities and limitations of different prediction approaches, ultimately promoting the adoption of user-friendly, data-driven solutions that empower farmers to enhance crop productivity and sustainability.

References

[1] World Bank, “Employment in agriculture (\% of total employment) – India,” World Development Indicators, 2023.

[2] Sharma, R. (2023). Indian Agriculture – Status, Importance and Role in Indian Economy. International Journal of Agriculture and Economic Development, 12(2), 45–52.

[3] Singh, "Economic Survey 2024: Erratic Monsoon, Climate Change, Crop Diseases Double Food Inflation in 3 Years," Down to Earth, Feb. 2024.

[4] Ministry of Agriculture \& Farmers Welfare, "Final Estimates of Production of Major Agricultural Crops" for the 2023-24 Season, Press Information Bureau, 2024.

[5] World Bank, Employment in agriculture (\% of total employment) - India, 2022.

[6] P. Venkatesh, N. M. L. Nithyashree, V. Sangeetha, and S. Pal, "Trends in agriculture, non-farm sector and rural employment in India: An insight from state level analysis," Indian Journal of Agricultural Sciences, vol. 85, no. 1, pp. 42–49, 2015.

[7] P. Kumar, A. Choudhary, P. K. Joshi, R. Prasad, and S. K. Singh, "Multiple crop yield estimation and forecasting using MERRA-2 model, satellite-gauge and MODIS satellite data by time series and regression modelling approach," Geocarto International, vol. 38, no. 1, pp. 243–263, 2022.

[8] M. N. Prof., S. S. Bharatkumar, R. B. Spoorthi, N. J. Chinmayi, and C. Tejashwini, “Research Paper on Maximizing Crop Forecast Accuracy Through the Analysis of Soil Composition and Weather Patterns,” Int. J. Res. Publ. Rev., vol. 5, May 2024.

[9] N. S. Shuaibu, G. N. Obunadike, and B. A. Jamilu, “Crop yield prediction using selected machine learning algorithms,” FUDMA Journal of Sciences, vol. 8, no. 1, Mar. 2024.

[10] Prasad, Moturi Ajay, et al. "Crop Yield Prediction Using Machine Learning." International Journal of Engineering Research and Science \& Technology 20.1 (2024): 203-206.

[11] Pradhan, Komal. "Intelligent Farming Systems in Future Using Machine Learning: A Focus on India." International Journal For Science Technology And Engineering, vol. 10, no. 5, July 2024

[12] Shahab, Hammad, et al. "IoT-Driven Smart Agricultural Technology for Real-Time Soil and Crop Optimization." Smart Agricultural Technology (2025): 100847.

[13] Eccim, Đurić Olivera, et al. "Application of machine learning in agriculture." Agricultural technique 49.4 (2024): 108-125.

[14] R. K. Dhanaraj, K. Rajkumar, and U. Hariharan, "Enterprise IoT Modeling: Supervised, Unsupervised, and Reinforcement Learning," in Enterprise IoT, Springer, 2020, pp. 35–60. doi: 10.1007/978-3-030-44407-5_3

[15] R. Goswami et al., "Whither digital agriculture in India?" Crop \& Pasture Science, vol. 74, no. 3, pp. 179–194, 2023. doi: 10.1071/CP21624

[16] K. Pradhan, "Intelligent Farming Systems in Future Using Machine Learning: A Focus on India," International Journal For Science Technology And Engineering, vol. 10, no. 5, Jul. 2024. doi: 10.22214/ijraset.2024.63551

[17] Majumdar, Jharna, Sneha Naraseeyappa, and Shilpa Ankalaki. "Analysis of agriculture data using data mining techniques: application of Big Data." Journal of Big Data 4.20 (2017).

[18] Thombare, Ruchita, Shreya Bhosale, Prasanna Dhemey, and Anagha Chaudhari. "Crop Yield Prediction Using Big Data Analytics." International Journal of Computer \& Mathematical Sciences, Vol. 6, Issue 11, pp. 53-61, November 2017.

[19] Aghighi, Hossein, et al. "Machine learning regression techniques for the silage maize yield prediction using time-series images of Landsat 8 OLI." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11.12 (2018): 4563-4577.

[20] Terliksiz, A. S., and D. T. Altylar. "Use of deep neural networks for crop yield prediction: A case study of soybean yield in Lauderdale County, Alabama, USA." 2019 8th International Conference on Agro-Geoinformatics. IEEE, 2019.

[21] Rathnayake, R. M. S. M., et al. "Predictive Analytics Platform for Organic Cultivation Management." 2019 International Conference on Advancements in Computing (ICAC). IEEE, 2019.

[22] Guruprasad, Ranjini B., Kumar Saurav, and Sukanya Randhawa. "Machine learning methodologies for paddy yield estimation in India: a case study." IGARSS 2019. IEEE, 2019.

[23] Araby, Alaa Adel, et al. "Smart IoT monitoring system for agriculture with predictive analysis." 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST). IEEE, 2019.

[24] Randhawa, Sukanya, Jitendra Singh, and Jagabondhu Hazra. "Towards a ML based global crop identification model using limited SAR data-that is scalable across data-sparse geographies." IGARSS 2019. IEEE, 2019.

[25] Kumar, Y. Jeevan Nagendra, et al. "Supervised machine learning approach for crop yield prediction in agriculture sector." 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2020.

[26] Nikhil, R., B. S. Anisha, and Ramakanth Kumar. "Real-time monitoring of agricultural land with crop prediction and animal intrusion prevention using internet of things and machine learning at edge." 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). IEEE, 2020.

[27] Nalwanga, Rosemary, et al. "Design of an embedded machine learning based system for an environmental-friendly crop prediction using a sustainable soil fertility management." 2021 IEEE 19th Student Conference on Research and Development (SCOReD). IEEE, 2021.

[28] Tahaseen, Munavvara, and Nageswara Rao Moparthi. "An assessment of the machine learning algorithms used in agriculture." 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2021.

[29] Anandhan, K., et al. "Comprehensive study: machine learning \& deep learning algorithms for paddy crops." 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO). IEEE, 2021.

[30] Muthoni, Francis, et al. "Machine learning model accurately predict maize grain yields in conservation agriculture systems in Southern Africa." 2021 9th International Conference on Agro-Geoinformatics. IEEE, 2021.

[31] Singh, Dushyant Kumar, and Rajeev Sobti. "Role of internet of things and machine learning in precision agriculture: a short review." 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC). IEEE, 2021.

[32] Mariammal, G., et al. "Prediction of land suitability for crop cultivation based on soil and environmental characteristics using modified recursive feature elimination technique with various classifiers." IEEE Transactions on Computational Social Systems 8.5 (2021): 1132–1142.

[33] Kumar Gajula, Akshay, et al. "Prediction of crop and yield in agriculture using machine learning technique." 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2021.

[34] Sharma, Ankita, et al. "Early Prediction of Crop Yield in India using Machine Learning." 2022 IEEE Region 10 Symposium (TENSYMP). IEEE, 2022.

[35] Thirumal, S., and R. Latha. "Automated Rice Crop Yield Prediction using Sine Cosine Algorithm with Weighted Regularized Extreme Learning Machine." 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2023.

[36] Deshmukh, Tanvi, Anand Singh Rajawat, and Amol Potgantwar. "Machine Learning Technique for Crop Selection and Prediction of Crop Cultivation." 2023 International Conference on Advanced Computing Technologies and Applications (ICACTA). IEEE, 2023.

[37] Kumar, T. Sathies, et al. "Crop Selection and Cultivation using Machine Learning." 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS). IEEE, 2023.

[38] Pawar, Rajendra, et al. "Crop Advancement with Machine Learning." 2023 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA). IEEE, 2023.

[39] Sunil, G. L., V. Nagaveni, and U. Shruthi. "A review on prediction of crop yield using machine learning techniques." 2022 IEEE Region 10 Symposium (TENSYMP). IEEE, 2022.

[41] Vanitha, K., G. Surya, and M. Rashmi. "Enhancing Predictive Accuracy for Agricultural Crop Yields in Indian States Using Power Transformation in Machine Learning Models." 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2024.