ISSN: 2093-5374 / E-ISSN: 2093-5382

Integrating Deep Reinforcement Learning and Feature Selection for Improved Crop Yield Prediction

A.K. Saritha[1*], and Dr. I. Jeena Jacob[2]

[1*]Department of Computer Science and Engineering, GITAM (Deemed to be University), Doddaballapur, Bengaluru, India. aksaritha@gmail.com, https://orcid.org/0000-0002-7031-238X

[2]Professor, Research Supervisor, Department of Computer Science and Engineering, GITAM (Deemed to be University), Doddaballapur, Bengaluru, India. ijacob@gitam.edu, https://orcid.org/0000-0001-6706-1017

Received: September 03, 2025; Revised: October 17, 2025; Accepted: November 18, 2025; Published: December 15, 2025

Abstract

One prominent area of research focuses on predicting crop production using parameters related to crop characteristics, soil properties, water availability, and environmental conditions. Critical crop characteristics are often gathered for prediction using machine learning algorithms. Although these methods show promise in tackling the yield prediction challenge, they still possess certain limitations and drawbacks. It is unable to map the initial information as well as crop yield figures directly, either non-linearly or linearly; the quantity of the attributes obtained has a significant impact on such algorithms' performance. Deep reinforcement learning encourages new aspects and dimensions for improving the previously listed shortcomings. Deep reinforcement learning creates a comprehensive framework for predicting crop production by fusing the smarts of reinforcement training with deep learning, which enables the mapping of unprocessed information to crop [2]forecasting values. In this work, we suggested a Deep Q-Network model for predicting various crop yields. Further, feature selection methods such as PCA, SelectKBest, and Random Forest Methods for feature selection. Based on that, rice, wheat, maize, cotton, and soyabean crops yield is predicted. In our comparative analysis, the Deep Q-Network model demonstrated a significant improvement over traditional methods, achieving an accuracy of 92% in predicting rice yield as opposed to the Random Forest model's 90% accuracy. The model outperformed the RF model in terms of sensitivity and specificity, both of which scored greater than 85%.

Keywords: Crop Yield Prediction, Deep Reinforcement Learning, PCA, Select K Best, Random Forest, and Feature Selection.

1 Introduction

Crop production and successful development are vital aspects of our everyday existence. Farmers may enhance their environmental management practices and create fairer, marketable products at a reduced cost by developing models that will boost crop model efficiency (Burdett & Wellen, 2022). In many nations, including India, agriculture is essential to the development of the national economy. Therefore, forecasting crop output in advance of cultivation may assist agricultural departments and farmers in

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 16, number: 4 (December), pp. 490-521. DOI: *10.58346/JOWUA.2025.I4.028*

*Corresponding author: Department of Computer Science and Engineering, GITAM (Deemed to be University), Doddaballapur, Bengaluru, India.

avoiding misinformation regarding selecting the best crop to plant and in implementing the necessary marketing and storage strategies. Prediction is one of the applications of deep learning of agricultural production (Shetty et al., 2021). Thus far, a large number of models have been approved and authorized (Elavarasan & Vincent, 2020). Because crop production forecast depends on several factors, like climate, soil, setting, area, period, and so forth, it requires the usage of certain datasets (Gong et al., 2022). Digitalization is becoming more and more important these days in a variety of knowledge fields, including weather forecasting, Internet of Things (IoT), agriculture, medicine, and recommendation platforms. Agriculture relies on crop yield estimation to increase productivity and support decision-making, particularly financial market prediction and encounters, and concerns related to food security (Moraye et al., 2021). From the dawn of the century, it's evident that the use of machine learning in agronomy has grown rapidly (Alibekova et al., 2025). This includes data-driven crop production projections derived from farm-level data regarding soil, climate, and management. The impact of data partitioning strategies on the models' actual performance, particularly when they are developed for yield forecasting, is, however, little understood (Nyéki & Neményi, 2022). Precision farming and food safety continue to aspire to quick and precise yield estimations, given the expanding range and availability of worldwide satellite products, as well as the quick development of new algorithms. It is still necessary to investigate the consistency and dependability of appropriate approaches that provide precise crop yield results (Liu et al., 2022). Decision-making in the food industry is largely dependent on precise crop output projections. The disadvantage is that a single-size-fits-all prediction model is often used for a whole region without taking into consideration regional differences in subregional VCI and TCI caused by environmental and climatic factors (Manoj et al., 2022).

Information on the anticipated crop output for this year's growing season is useful for farmers, policymakers, and food processing facilities. Increasing crop revenue is one of the key advantages of using trustworthy forecasting technologies (Mousa & Rasheed, 2022). Selecting the right agricultural product management approach is aided by knowledge about the quantity of crop produced prior to harvest. The right choice of independent variables affects how challenging it is to create forecasting models (Elavarasan & Vincent, 2020). In India, agriculture is the most common employment and provides survival for around half of the people. Given that a variety of elements, such as the environment, precipitation, soil, water, and seasonal crops, affect crop output prediction, it is a popular area of study. The extraction of significant crop traits for yield prediction is a popular use of ML models. The CYP uses the ML models as an aid to decision-making to determine what kinds of crops may be grown and when (Shook et al., 2021; Abbasi et al., 2025). Together with water and clothing, food is one of the most important needs for human life. In India's GDP, agriculture is the primary industry, followed by the IT sector. Choosing the right crops at the right time of year is essential to achieving a high crop output. Farmers often use techniques that are based on popular belief and ancestry, both of which have been shown to be ineffective (Elavarasan & Vincent, 2021). In terms of contribution to the Indian nation's economic output, agriculture ranks third. Choosing the appropriate crop for the correct soils at the right time of year is crucial to achieving a high crop yield. The great majority of farmers rely on their gut feelings when choosing which season to plant their crops, even if doing so might result in losses. Important judgments, such as predicting agricultural output for a certain kind of soil or season, may be made with algorithms based on ML to separate vast amounts of data (Bali & Singla, 2021). Many nations throughout the globe rely heavily on agriculture, and predicting crop production is essential to ensuring food security. The techniques for calculating the yield include combining several data sources, such as field data and meteorological data, with temporal data. Creating a machine learning-based multi-resource data-based estimation model is one of the main issues. Numerous studies have been conducted from planting to harvest (Josephine et al., 2020). In order to address the impending dilemma, emerging

nations like India should concentrate on implementing novel agricultural technology due to their rapid population expansion. Early crop production prediction is a crucial problem in precision farming as it requires a comprehensive knowledge of the development pattern with highly nonlinear characteristics, making it one of the most difficult tasks. Field-to-field variations exist in the dynamic approach of management methods such as fertilizers, herbicides, and irrigation, as well as environmental elements including temperature, humidity, and rainfall (Sivanandhini & Prakash, 2020). Even while it still accounts for the majority of revenue in many nations, food insecurity brought on by climate change has had a detrimental effect on the agricultural sector in recent years. This is because most crops suffer from harsh weather conditions brought on by climate change, which also adversely affects the expected level of agricultural outcomes. Unfortunately, there is no proven methodology to completely prevent these natural occurrences. It would be preferable if farmers had advanced knowledge of what lies ahead so they could make appropriate plans. Risk reduction for food poverty may be aided by early information exchange about anticipated crop output (Sunil et al., 2022). Cultivating plants and animals is the science, along with the art of agriculture. India emerges second in the world for farming, which uses 60.45% of the country's land. The agricultural sector and farming are the main drivers of the Indian economy (Monica Nandini, 2024). Agriculture benefits greatly from a variety of conditions, including rainfall, air and surface temperatures, soil moisture, crop rotation, and soil elements like potassium, phosphorus, and nitrogen (Ravi & Baranidharan, 2020). Farmers can make better decisions about when and what types of crops to sow by using a crop yield prediction system, depending on environmental conditions, in order to maximize production. The purpose of the advanced ensemble regression crop yield forecasting model is to forecast crop production by taking into account several phenotypic parameters, such as solar radiation, precipitation, maximum and minimum temperatures, etc. The dataset pertaining to maize and soybean crops has a 38-year history of yield performance information gathered from 105 distinct sites. A comparison based on feature selection between mutual information and correlation. Most associated characteristics with crop output and improved crop yield forecasts via mutual information (Muruganantham et al., 2022).

The paper's remaining sections are arranged as follows: the related study on the work covered in Section 2. You may find the proposal's research approach in Sections 3 and 4. Findings from experiments are discussed. The paper ends with Section. 5, which integrates an overview of the direction crop yield prediction improvement is headed.

2 Related Work

Authors created deep learning-based models in to conclude the efficiency of the edge cutting algorithms in relation to several performance metrics. The machine learning method XGBoost, convolution neural networks, XGBoost, along with recurrent neural networks are the algorithms they analyzed in this research. Using ecological, soil, silt, oxygen, clay, ocd (Organic Carbon Density), the OCS (Organic To predict crop production, the case study included the following factors: carbon stock, pHH2O (soil pH in water), sandy soil, soc (soil organic carbon), ceo (cation exchange order), water, and crop factors. Their suggested approach performs very well by using feature selection to forecast crop production. Their suggested technique for forecasting yields was applied to the corn and soy crops. The study (Anbananthen et al., 2021) proposes hybrid machine learning algorithms that employ certain ensemble techniques as gradient boosting, random forest, selection operator regression, and least absolute shrinkage mixed with layered generalization, gradient boosting, random forest, and selection operator regression. A new model known as stacked generalization finds the best way to combine predictions from several models that have been trained on the same dataset. The new method is demonstrated using

the aerial-intel dataset from the GitHub data analysis project. The following conclusions have been drawn from the experimental findings on the agricultural data the effectiveness of the individual and hybrid machine learning methods is compared using cross-validation to ascertain which ones work best for the farming dataset. Accuracy rates for the random forest regressor, stacking generalization ensemble, and gradient improved tree regression techniques are 88.71%, 86.98%, and 88.89%, respectively. With a statistically significant accuracy of 88.89%, the proposed stacking generalization machine learning algorithm performs better than the others, suggesting that it is an effective crop production prediction algorithm. The farmers get prompt and precise feedback from the system as well. The primary goal of the paper in (Anbananthen et al., 2021) is to use hybrid methods of machine learning to anticipate agricultural production more accurately. Using specialized ensemble techniques, including layered generalization, gradient boost, and random forest, with LASSO regression, this paper suggests hybrid machine learning algorithms. A new model known as stacked generalization finds the best way to combine predictions from several models that have been trained on the same dataset. In order to illustrate whether the proposed method may be working successfully, the aerial-intel dataset from the GitHub information science repository is used. The following conclusions have been drawn from the experimental findings on the agricultural data. The competitiveness of the individual and hybrid machine learning methods is compared using cross-validation to ascertain which ones work best for the farming dataset. With a statistical outperformance of 88.89%, the suggested stacking generalization machine learning method shows that it is a useful technique for crop production prediction. The farmers get prompt and precise feedback from the system as well. In (Oikonomidis et al., 2022), scientists developed deep learning-based models to assess the performance metrics of the underlying algorithms and ascertain their level of success. The methods that they examined in this study included CNN-XGBoost, CNN-Long Short-Term Memory, CNN-Deep Neural Systems, CNN-XGBoost machine learning algorithm, and CNN-Recurrent neural network training. Using 25,345 samples and 395 characteristics such as weather as well as soil conditions from a publicly available soybean dataset, researchers ran experiments for the case study. The RMSE of 0.266, MSE of 0.071, and MAE of 0.199 of the hybrid CNN-DNN system demonstrated superior performance compared to the other models. The model has an R2 of 0.87, meaning it predicts the data. Compared to the other DL-based models, the XGBoost method took less time to execute and produced the second-best result. The authors of (Sajid et al., 2022) broaden the idea to include the whole US Corn Belt (12 states). More precisely, taking into account the situations regardless of crop modeling variables, they constructed five novel machine learning models and associated ensemble models. They also use soil, weather, and management, along with historical yield data as more input values in their models. Their study is distinct in that they employ spatial analysis to look at the reasons behind high or low model prediction mistakes. Their findings showed that combining crop modeling and machine learning improves forecast accuracy. The ensemble model performed better than the individual machine learning models, with an average root-mean-square error of about nine percent for the three test years (2018, 2019,2020), which is in line with previous research. Furthermore, counties with crop reporting districts that have low cropland ratios also have high RRMSE, according to an examination of the causes of inaccuracy. Moreover, they discovered that large inaccuracies in some places were caused by severe weather occurrences and the soil input data. If data is available, the suggested models may be used for field-level prediction as well as large-scale forecasting at the county level. A strategy to handle the two problems and increase the accuracy of crop yield prediction is presented in the study (Pham et al., 2022). This is accomplished through the use of (i) higher-order spatial autonomous component analysis and (ii) a combination of machine learning and principal component analysis (PCA). Utilizing Vietnam as an example, the suggested framework combines common VCI/TCI distributions into each of their subregions. Subregional crop rice forecasting systems across Vietnam outperformed the one-fits-all method by an average of 20% to 60%. PCA-ML combo

surpassed ML-only by as much as 45% on average. The framework demonstrates its dependability by producing rice production projections with an average inaccuracy of 5% up to two months before the harvest. The article (Hara et al., 2021) presents and discusses the most widely used independent factors in agricultural crop production prediction modelling using neural network models. Particular attention is paid to environmental elements such as air temperature, total rainfall, insolation, and soil properties.

Authors examine in detail the potential applications of vegetation indices and plant production indices, which are useful indicators derived from the use of remote sensing methods. The study emphasizes how the increasing use of remote sensing and photogrammetric methods enables precision agriculture. Furthermore, some constraints regarding the utilization of particular input variables are delineated, along with additional prospects for the advancement of non-linear modeling via the deployment of artificial neural networks, being a mechanism that bolsters the pragmatic implementation and enhancement of precision farming methodologies. The goal of the (Park et al., 2023) research is to use meteorological factors that are strongly connected to crop development to construct a crop production forecasting model at wide regional scales. Local soil, along with environmental variables, might result in a geographically heterogeneous impact of climatic patterns on agricultural output (Geetha, 2024). In order to forecast county-level maize output for 5 Midwestern states, they suggest using a Bayesian spatially variable functional model. The daily maximum and lowest temperature trajectories, as well as the annual precipitation, are used as multivariate functional predictors in this model. By allowing functional coefficients to fluctuate across space, the suggested model respects the geographically diverse connection between the response and associated predictors while taking into consideration spatial correlation and functional predictor measurement error.

To further increase its ability to handle geographic heterogeneity, the model also includes a Bayesian choice of variables mechanism. The suggested approach is shown to perform better in predicting corn output than other very competitive techniques because it is flexible enough to accommodate geographic heterogeneity with spatially variable coefficients in its model. Their research adds to their knowledge of how agricultural productivity is impacted by climate change. Additional resources for this subject may be found on the internet. The study (Pham et al., 2022) established a baseline of non-feature reductions (All-F) and provided a technique for analysing the effectiveness of FS, FX, and FSX together.

21 ML-based rice yield forecasting algorithms linear. In Vietnam, the vegetation scenario index (VCI)/temperature situation index is used to generate support vector machines, decision trees, artificial neural networks, and ensembles across eight sub-regions. The findings show that FSX fully utilizes the FS and FX, which leads to FSX-based models outperforming FS-based models in two (1) Fg-based (FX-based) designs and 18 out of 21 types. These FXS-, FS-, and FX-based models outperform all-F-based models at a mean level of 21% and 60% in terms of RMSE. Furthermore, the Ensemble (13 models) served as the foundation for the creation of 21 of the best models.

Tree (6 models), linear (1 type), and ANN (1 method) frameworks. These results demonstrate the important role that FS, FX, and particularly FSX play when combined with a variety of ML algorithms (particularly Ensemble) to improve crop yield prediction accuracy. The suggested approach in (Sajid et al., 2022) seeks to provide a quick and accurate way to anticipate agricultural output rates while also offering a substitute crop based on specified parameters. Therefore, the approach predicts crop yield rate and alternative crop yield rate by applying machine learning methods and algorithms like Random Forest Regressor and Linear Regression to discover patterns among information and process it according to the input condition. A crop production forecast model that incorporates the results of an APSIM crop simulation model was presented in (Shafi et al., 2023). To increase transparency, these results are included as characteristics in ML models. An ensemble model is then created for prediction. The maize

yields of the twelve US maize Belt States are predicted using the suggested model. After assessing the framework at the level of the county for the years 2019, 2020, and 2018, respectively, RRMSE of 9.34%, 8.99%, along 9.52% was attained. The suggested model is interpretable and performs as well as the best-performing state-of-the-art prediction models in terms of the performance metrics. Further investigations have been carried out to comprehend the origin of model faults. Because there is more data available in counties with crop reports districts that have substantial cropland ratios, the study has shown that these locations have lower error rates. Researchers in (Umamaheswari, 2025) propose that, in order to train yield models for prediction on a horizontally networked dataset spread across multiple client devices, federated learning is used. In particular, regression models based on deep residual networks, such ResNet-16 and ResNet-28, are trained using the federated averaging methodology in order to forecast soybean output in a decentralized setting. These models' performance is then contrasted with that of deep learning and machine learning techniques. When compared to centralized learning models, federated averaging using the ResNet-16 regression framework with the Adam optimizer yielded the greatest results, according to the findings of experimental learning models. This method is also simply implementable for the prediction of yield in a federated context. A Support Vector Machine, or SVM, is used in (Sadasivam, 2021) work to forecast agricultural output under various environmental parameters, including soil, climate, and biological components. Granular computing allows breaking down the issue area into a series of smaller jobs. Thus, partitioning the problem area allows for the parallelization of the SVM hyperplane building. Another option is to parallelize testing. The primary benefit is that a greater dimension space may be handled by a linear SVM. There is less complication in time. Granular SVM prediction may be parallelized using suitable methods such as MapReduce/GPGPU. Crop output is increased via IoT-based agriculture through precise forecasting, automation, remote surveillance, and less resource waste. Farmers, academics, and government officials may employ IoT-based monitoring systems to evaluate statistical data and agricultural environments to predict crop yields.

In this research, a distributed granular support vector machine-based Internet of Things system for predicting crop yield based on biological, soil, and meteorological parameters is proposed. Two architectures have been suggested in the (Nejad et al., 2022) research. 2D-CNN, connections that skip, and LSTM-Attentions are included in the initial model. The second model consists of 3D-CNN, ConvLSTM Attention, and skip connections. For 1800 counties in the United States, the MODIS products Land-Cover, Surface-Temperature, and MODIS-Land-surface offer county-level input data, where soybeans are mostly grown, from 2003 to 2018. The most current models and the suggested techniques have been contrasted. Then, the outcomes demonstrated that the second suggested way worked noticeably better than the previous approaches. Regarding MAE, the subsequent suggested approaches, Deep Yield, ConvLSTM, 3DCNN, as well as CNN-LSTM, yielded results of 4.3, 6.003, 6.05, 6.3, as well as 7.002, in that order. In the framework of (Shafi et al., 2023), an approach for estimating wheat grain output using three regression methods is presented: random forests, Extreme Gradient Boosting regression, and Least Absolute Shrinkage & Selection Operation (LASSO) regression. Numerous factors of each of the above approaches are analyzed, and the results are comparatively studied, in order to identify the optimized approach. Data from three wheat field trials with three distinct sowing dates (SD1, SD2, and SD3) are collected using multispectral drone-based sensors. After that, the consequence of the seeding plan on crop yield is examined (Haghighi & Far, 2014). The models' forecasting ability is assessed using a range of evaluation criteria at different stages of the crop's growth. With an absolute mean of 21.72 and a coefficient of estimation (R2) of 0.93, the data indicate that LASSO performed at its best in April. The average annual projected yield of the wheat cultivated in the fields with SD1, SD2, and SD3 was 260.54 g/m2, 201.64 g/m2, and 47.29 g/m2,

respectively. This research may assist agronomists and farmers in making well-informed choices on resource management, planting and harvest schedules, and other aspects of crop management.

3 Proposed Work

1) Research Problems

The suggested methods for predicting crop production are presented in this section Figure 1. Forecasting the yield of crops is used to forecast production in the agricultural industry in order to improve the cultivation process, as well as make long-term strategic decisions that would increase crop yield. The prediction model that has been suggested can be transformed into a decision support system (DSS) for precision agriculture, which strives for whole farm management.

The enormous burden of applying AI approaches to agricultural production forecasting is made more difficult by the lack of a uniform model, which makes algorithm creation difficult. Regression algorithms, as well as neural networks, show the greatest promise for predicting agricultural production. Benefits of machine learning for predicting crop production:

1. Complex dataset – Predicting agricultural yields requires analyzing large datasets made up of historical and/or satellite imagery. Algorithms for regression (SVR, RF) as well as neural networks (CNN) are two examples of AI approaches that may be used to make forecasts more quickly and accurately.
2. Parameter variation – Numerous variables, including climate, the state of the soil, NDVI, altitude, and air parameters, affect crop output. The AI-powered prediction systems effectively manage the dependence of variables.
3. Accurate prediction – Low error indexes, such as RMSE as well as R2, which are common indicators of correctness for statistical evaluation, are displayed by parameter predictions using ML.

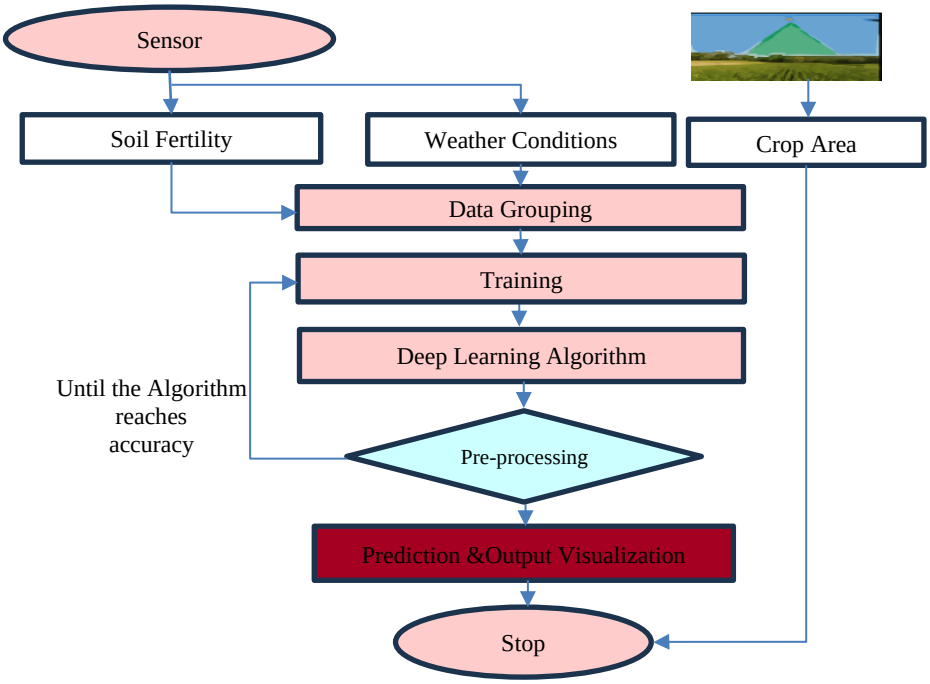


Figure 1: Proposed system flowchart

Obstacles and constraints in agricultural yield forecast:

1. Complex datasets, as well as changing factors, make it difficult for prediction algorithms to be universally designed.
2. Since of this intricacy, choosing the right database is essential since a poor choice of data might lead to an unfit or overfit projection pattern.
3. A lack of a common set of features for yield prediction and analyses over a set of sub-features, which affects the yield prediction and underlying aspects that vary for the crop in variance with climate and other vegetational environments.

2) Preprocessing

Processing includes tasks such as discretization, alteration, and cleansing of the information. The procedure of filling in missing numbers, eliminating noise from data, and fixing any discrepancies in the data is known as data cleaning. Normalizing data to make it suitable for data analysis is known as transforming it. The process of translating continuous values into an array of attribute intervals is known as information discretization. Figure 2a illustrates the simulation process used for crop yield prediction. In this work, for example, we employed the technique of breaking down the factor that is dependent, crop yield, into three separate values: Low Yield, Average Yield, and High Yield. The Deep Q-Network (DQN) environment was modeled using normalized features such as weather conditions, soil humidity, and soil fertility, where each state represented a feature vector and the action corresponded to a predicted yield class Figure 2b. The agent received a reward of +1 for correct predictions and 0 or –0.1 otherwise, with each episode consisting of 100 randomized samples. The Q-network comprised two optional 1D convolutional layers followed by three intermediate layers with 128, 64, and 32 neurons activated through the ReLU function, ending with an output layer that produces Q-value predictions for various yield categories. The network was optimized using the Adam algorithm with a learning rate of 0.001, a discount factor (γ) of 0.95, and an ε-greedy exploration policy where ε decayed from 1.0 to 0.1. Training was carried out with a batch size of 32 over 500 episodes per crop, leading to consistent and stable model convergence.

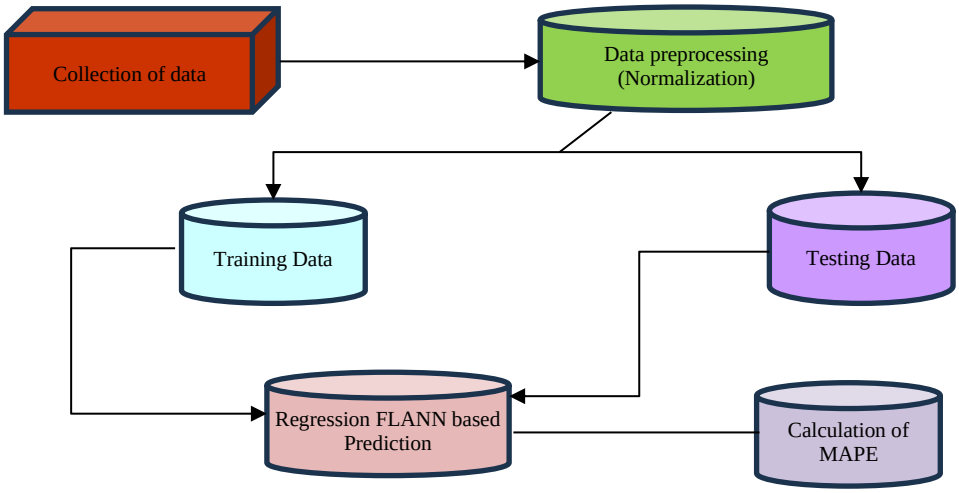


Figure 2a: Simulation diagram

Column	Non-Null Count	Dtype
Year	3400	int64
Dist Name	3400	object
Rainfall (mm)	3400	float64
Temperature_Max (Celsius)	3400	float64
Temperature_Min (Celsius)	3400	float64
RICE AREA (1000 ha)	3400	float64
RICE PRODUCTION (1000 tons)	3400	float64
WHEAT AREA (1000 ha)	3400	float64
WHEAT PRODUCTION (1000 tons)	3400	float64
MAIZE AREA (1000 ha)	3400	float64
MAIZE PRODUCTION (1000 tons)	3400	float64
SOYABEAN AREA (1000 ha)	3400	float64
SOYABEAN PRODUCTION (1000 tons)	3400	float64
COTTON AREA (1000 ha)	3400	float64
COTTON PRODUCTION (1000 tons)	3400	float64
Yield (hg/ha)	3400	float64
NITROGEN CONSUMPTION (tons)	3400	float64
NITROGEN SHARE IN NPK (Percent)	3400	float64
NITROGEN PER HA OF NCA (Kg per ha)	3400	float64
NITROGEN PER HA OF GCA (Kg per ha)	3400	float64
PHOSPHATE CONSUMPTION (tons)	3400	float64
PHOSPHATE SHARE IN NPK (Percent)	3400	float64
PHOSPHATE PER HA OF NCA (Kg per ha)	3400	float64
PHOSPHATE PER HA OF GCA (Kg per ha)	3400	float64
POTASH CONSUMPTION (tons)	3400	float64
POTASH SHARE IN NPK (Percent)	3400	float64
POTASH PER HA OF NCA (Kg per ha)	3400	float64
POTASH PER HA OF GCA (Kg per ha)	3400	float64

Column	Non-Null Count	Dtype
TOTAL CONSUMPTION (tons)	3400	float64
TOTAL PER HA OF NCA (Kg per ha)	3400	float64
TOTAL PER HA OF GCA (Kg per ha)	3400	float64
CANALS AREA (1000 ha)_x	3400	float64
TANKS AREA (1000 ha)_x	3400	float64
TUBE WELLS AREA (1000 ha)_x	3400	float64
OTHER WELLS AREA (1000 ha)_x	3400	float64
TOTAL WELLS AREA (1000 ha)_x	3400	float64
OTHER SOURCES AREA (1000 ha)_x	3400	float64
NET AREA (1000 ha)_x	3400	float64
GROSS AREA (1000 ha)_x	3400	float64
CANALS AREA (1000 ha)_y	3400	float64
TANKS AREA (1000 ha)_y	3400	float64
TUBE WELLS AREA (1000 ha)_y	3400	float64
OTHER WELLS AREA (1000 ha)_y	3400	float64
TOTAL WELLS AREA (1000 ha)_y	3400	float64
OTHER SOURCES AREA (1000 ha)_y	3400	float64
NET AREA (1000 ha)_y	3400	float64
GROSS AREA (1000 ha)_y	3400	float64
District Code	3400	float64
Boron (ppm)	3400	float64
Zinc (ppm)	3400	float64
Sulphur (ppm)	3400	float64
Potassium (ppm)	3400	float64
Phosphorus (ppm)	3400	float64
Organic Carbon (ppm)	3400	float64
Electrical Conductivity (dS/m)	3400	float64

Figure 2b: Dataset description

3) Feature Selection

A crucial technique for reducing the dimension of data that is frequently applied in machine learning algorithms is feature selection. The feature selection techniques based on a limited representation are particularly appealing in this context. This is due to the way these techniques work, which seeks to represent information with the fewest feasible non-zero components.

Network data typically contains a large number of elements that are either irrelevant or don't yield a lot of information. These parameters have no discernible effect on the categorization findings. The feature selection procedure largely eliminates these superfluous attributes, which also lowers their detrimental consequences on the rating system.

A very high-dimensional space for features is typically present in deep neural networks. One might consider the benefits of feature selection procedures in such a scenario. This research initially compares three minimal choices of feature techniques: PCA, Random Forest, and Select Best Figure 3. The feature evaluation model is being used. Information Gain utilizes entropy assessment at the second level, while Principal Component Analysis uses feature correlations at the first level.

A number of parts make up the design: i. primary Component Analysis: In this process, the data set's fundamental elements are produced. The most important elements are found by analyzing the association among the characteristics. With the use of a conversion method called PCA, it is possible to condense big data sets with a lot of interconnected characteristics so that the present data may be described using fewer variables. ii. Information gain (IG) evaluation of features: In this step, the already chosen feature set is assessed to find the most relevant characteristics. The last aspect set is chosen by calculating the IG for elements and using a predetermined threshold (t). iii. Model Training and Categorization: In this step, a training data set is used to train the classification algorithm or algorithms. Later, test data and fresh data sets are classified using the trained model Figure 4.

The Random Forest method (Random Forest) is an automatic learning methodology. It uses bootstrap, random resampling, and node stochastic classification technology to construct multiple decision trees and votes to determine the final category outcome. Complex categorization features may be analyzed using RF, which can also quickly learn new features and is resilient to missing and redundant data. The random forest algorithm's variable significance measure is a crucial component. For feature selection in this work, we use a permutation of the Area Under the Curve (AUC)-based variable importance measure (VIM) (Ravi & Baranidharan, 2020). Generally, it provides four methods for measuring the components' significance.

An AUC-driven permute. By computing the average decrease before and after the region under the line after the mild disruption of the information that is an independent variable, VIM primarily reveals the significance of the variable. Consequently, the variable receives an importance measure $VI_{x_j}^{AUC}$ of x_j predictor is calculated as follows:

Predictor equation (1):

$$VI_{x_j}^{AUC} = \frac{1}{ntree}\sum_{t=1}^{ntree^*}\left(AUC_{tj} - AUC_{tj^-}\right) \quad (1)$$

ntre e^* shows the forest's trees, AUC_{tj} (1) shows the AUC determined using OOB data prior to permuting predictor x_j in tree t, AUC_{tJ} shows the AUC derived from the OOB data after predictor permutation x_j randomly in tree t.

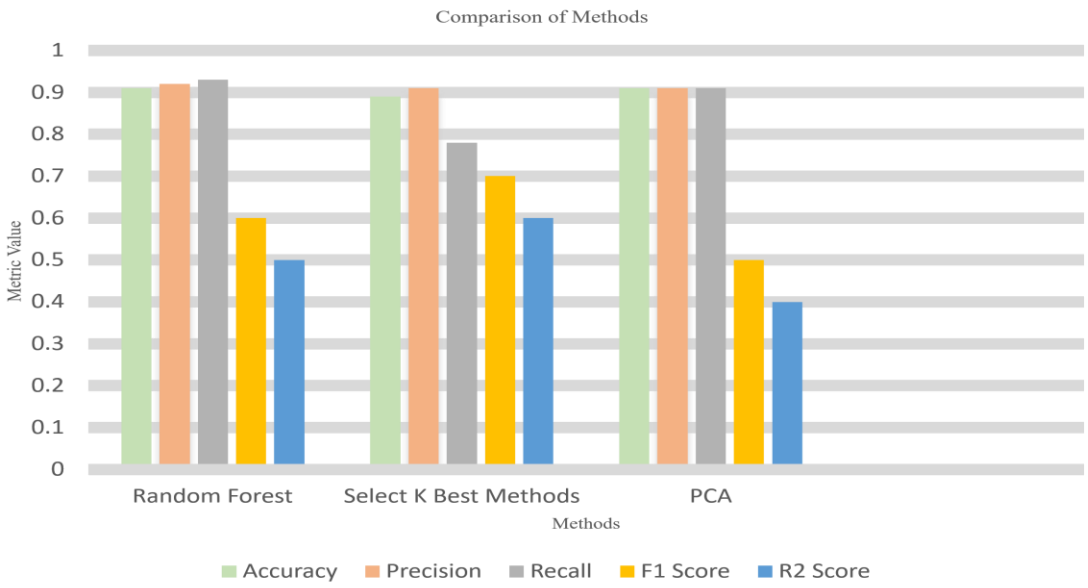


Figure 3: Comparison of methods

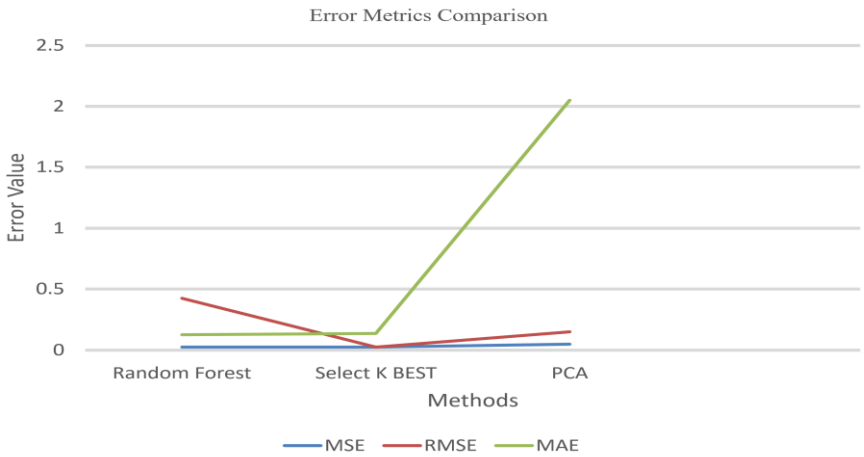


Figure 4: Error metrics comparison

4) DRL

With the rapid expansion of data and enhanced measurement persistence, deep reinforcement training has developed to create new options for identifying, assessing, and acknowledging large-scale data processes for agricultural frameworks. Among the crucial elements that must be examined in order to structure the deep reinforced models for learning are:

Recognizing the fundamental patterns and structures within the constrained sample space. The tasks associated with the continuous representation of events are being reviewed. The framework must function well enough to accommodate activities that are always evolving. The next part provides a detailed explanation of the deep Q-Network technology, Q-learning, and reinforcement learning. A yield prediction environment is constructed based on input parameters that converts the controlled learning process into a reinforcement learning procedure. One could characterize the setting as a "yield prediction

game." Each game includes a collection of samples and the labels that go with each combination of parametric features and thresholds that help improve crop productivity.

When an agent first begins playing the game, it estimates the values of the crop production parameters by completing the tasks to receive rewards. Every nearby anticipated quantity for the goal yields a positive reward for the agent, whereas every other location yields a negative benefit.

The representative will be given a final score for all of the activities they took during the procedure.

The Markov decision-making procedure (MDP), which is an equation made up of S, A, T, and R where S is a set of indicates, A is a set of decisions, T is the mapping function that allows every state-action pair to move to a new state—is the foundation of reinforcement learning, which is a method used in machine learning. R is the function of rewards that results from this process. T, which has the Markov property, controls every phase in the MDP from the present phase-action combination to a subsequent state. As a result, when the MDP is defined, its approach is a one-to-one translation of every state to an activity. This means that the MDP makes it possible to choose the appropriate course of action for every state in order to optimize the overall anticipated reward R. The approach, which is in accordance with the appraisal of what was done in the present situation Q, may be gathered, with V corresponding to the appraisal of every state, as well as Q corresponding to the appraisal of every state action pair. The optimum criterion is often connected through the worth operate V, and this is an estimation of the worth of every state. Utilizing the ε-greedy technique, investigate the actions that need to be carried out in the current state while keeping a close eye on the condition-action space in order to get the optimal model policy. With probability p, the agent will select the present state, whereas with likelihood 1 − p, it will select random actions. The agent matches the real V as well as Q functions by constantly communicating with the surroundings and modifying its own V as well as Q function. The Q function, the data entered are examined, sorted by feature significance utilizing three FS classifiers, and the best subset of characteristics is selected. Redundant features are eliminated from the dataset, leaving just the subset of chosen attributes. In parallel, the information set is modified using Mini-Batch, and the DRL model receives the recoded data samples. A neural network trained using characteristics will extract the data input for the DRL system. The classification algorithm will then undergo DQN training, and the algorithm will determine whether the supplied transport is in two classes.

In this paper, $X = (X_l, X_2, \dots, X_n)$ is an $n \times d$ matrix with samples of d-dimensional information. $X_{i,j}$ alludes to a matrix component, and X^j and X_i correspondingly allude to the row j($j = 1,2, \dots, d$) in addition to the i th column ($i = 1,2, \dots, n$) within the grid. The total number of classes is denoted by c in the label matrix Y, which is an n×c matrix. $e_n = [1 \dots 1] \in R^n$ is a vector where each member is equal to one. W is the matrix of d×c mappings.

Figure 5 shows the input parameters, such as soil, weather, crop area, and other factors affecting crop yield prediction. The model's input consists of Mini-Batch samples chosen based on features, from which feature values are retrieved subsequent to convolutional layers. The activation mechanism of every layer in the network that is completely connected is then assessed after the characteristics have been flattened and fed into the three-layer fully interconnected layer Figure 6a. In a fully connected network, the ReLU function acts as every layer's activating function, guaranteeing that all computed Q values are positive.

The Deep Q-Network (DQN) model's overall processing flow is depicted in the Figure 6b. The Experience Replay Memory (E) receives a set of input states that represent the current state of the environment. By storing previous experiences as state–action–reward–next state tuples, this memory component improves stability and speed by allowing the model to learn from a variety of samples rather than sequential observations. The Q-network function, F(), is then used to process the recorded

experiences and estimate the Q-values, or the anticipated future rewards for each potential course of action.

The goal output, Ytw, which represents the expected reward or value connected to the selected action, is finally produced by the network. The DQN can iteratively improve its decision-making approach and converge towards optimal policy learning by using this prediction to adjust the network parameters through backpropagation.

The DRL model's entire layer of connections is where the DQN method is mostly used, and the model computes the prediction. $\hat{}_{At}$ and $a\hat{}_{t+1}$ corresponding to s_t and s_{t+1} states, in that order; if the anticipated action (a hat t) and the state st right action at this point are identical, the reward is 1; if not, it is 0, and the value of the reward is found as r_t.

Remarkably, assuming that the information being studied is labeled, as well as the labels are independent, the algorithm's reward discount rate is set to 0.01 to get the best possible performance, as well as motivate the algorithm to concentrate on the current learning reward. In order for the action chosen by the model, as well as estimated by its Q function, to receive the largest anticipated total reward in the current state.

Matching the Q operation, which represents the maximum predicted reward that the surroundings might offer in a certain circumstance and activity, is the main goal of the DQN algorithm. The system's activity and condition define the Q function. Once we have Q(s, a), we may get the policy function. $Policy\ (s) = arg_a\ max(Q(s,a))$ is a state-dependent strategy functional that chooses the course of action that will optimize Q.

The fundamental procedure of the algorithm, known as DQN is shown in Figure 4, where the model is given an instance of the Mini-Batch specified in the preceding paragraph, and each iteration regenerates all of the Mini-Batches. The equivalent $Q(s_t, a)$ and $Q(s_{t+1}, a)$ are calculated using the current conditions of each state and $a\hat{}_t$. The highest Q value in the present state is then ascertained by feeding those, the main fitted Q function, to the Policies module as previously indicated.

$Q(s_t, a)$ is further processed by the Policy method to provide the highest Q value, and the next action (label) that will be tried is chosen using the ε-greedy In order to calculate the reward amount, an algorithm with the likelihood of success is supplied into the Reward section and compared with what was actually done (label) $Q(s_{t+1}, a)$ is further calculated using the Policy function.

The q_t, q_{t+1} are then calculated using the corresponding choice made at and a_{t+1}, and the rt computed via q_{t+1} and the reward function obtains the reward as $q_{ref} = r_t + \lambda * q_{t+1}$, If the incentive is set to a 0.01 deduction factor, indicating therefore there is no connection between s_t and s_{t+1}. Next, we will get q_t and q_{ref}. To update the parameters of the DQN model, compute the loss value, and backpropagate across the initial network.

The DRL approach is used in addition to the FS procedure to accurately identify and categorize attacks on networks (Anbananthen et al., 2021). The DRL is a prominent algorithm for machine learning that trains an agent to choose the best way to get a predicted return. The Markov Decision Process (MDP) provides a simple theoretical framework for handling the DRL issue. During the process of communication, the situation reacts to the activity, the agent analyzes the current environment's states and chooses a particular policy π, and the agent is fed new state observations and rewards. Thus, it is assumed that the beginning is the early condition. s_0, By using the Markov decision process (MDP), one might potentially $s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_n, a_n,\ r_n$.

Improving the policy for acting in a way that takes advantage of the anticipated profit is the agent's job. The quantity of the discounted incentives is returned in step t. $G_f = \sum_{k=0}^{\infty} \gamma^k r_{r+k+1}$, whereas $\gamma \in [O, 1]$ shows the discount rate that establishes the current worth of future benefits. Using the activity function of value to teach agents to solve MDP problems may be a crucial strategy in reinforcement learning $Q_\pi(s; a)$. $Q_\pi(s;$ a) describes the action's predicted return that is taken into consideration based on the state's policy πs.

Activity equation (2):

$$Q_\pi(s, a) = E_\pi[[G_t \mid s, a], \quad (2)$$

$Q_\pi(s, a)$ (2) determines the action value for a given state. Generally, $Q_\pi(s, a)$ mention the advantages of the agent being in a certain condition. As a result, the optimal strategy depends on the ideal quantity. Particularly after the ideal action value function $Q^*(s, a) = max_\pi Q_\pi(s, a)$ is reached, the ideal course of action $\pi^*(s) = argmax_{axA}\ Q^*(s, a)$ chooses the course of action that aligns with the maximum $Q^*(s, a)$ in all the states. Usually, $Q^*(s, a)$ is clarified as Bellman optimality, which shows the connections between the current suitable action function of values and the next optimal action values function. The next function, which is used to determine the next action value, is described as the Ideal value function.

Ideal action value equation (3):

$$Q^*(s, a) = E_{s'}[[r + \gamma max_{d'} Q^*(s', a') \mid s, a], \quad (3)$$

Whereas s' indicates the condition obtained subsequent to the activity conducted. A, and a' show the action taken into consideration in the subsequent state. After running through Eq.(2) many times, it finally converged to the ideal action value function $Q^*(s, a)$.

The simulation for this neural network model was conducted using Python with TensorFlow (Keras API) for neural architecture design and training. The dataset was pre-processed using scikit-learn, where features such as weather parameters, soil humidity, and soil fertility were normalized and used as inputs. The network architecture comprised an input layer representing the environmental parameters, a hidden layer with three neurons (F1, F2, and F3) using ReLU activation functions, and an output layer producing the final crop yield value. The model was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 32, and 100–200 epochs under early stopping to prevent overfitting. The data was divided into 70% training, 15% validation, and 15% testing sets.

Software Details

The primary programming language utilised was Python.

For the neural network model's implementation, with the deep learning architecture is designed and trained with the aid of the TensorFlow and Kera’s frameworks. NumPy and Pandas were used for numerical operations and dataset handling, while Scikit-learn was used for data preprocessing, feature normalization, and evaluation metric computation. A machine with an Intel Core i5 processor, 8–16 GB of RAM, and an optional GPU (NVIDIA CUDA-enabled) for quicker computing was used to run the simulations in a Windows 10 environment.

Yield Prediction

This focusses on predicting local yields by using the crop that is sown in the area. For some crops, production projections are established at the district level, like well as the best-yielding plants are

recommended for the future in order to maximize the analysis's benefits for producers. Using the dataset, a number of regression models were applied to predict production. The prediction of crop yield is challenging with lack of proper set of common features and complicated with complex datasets

Yield equation (4):

$$Yield = Production / Area \quad (4)$$

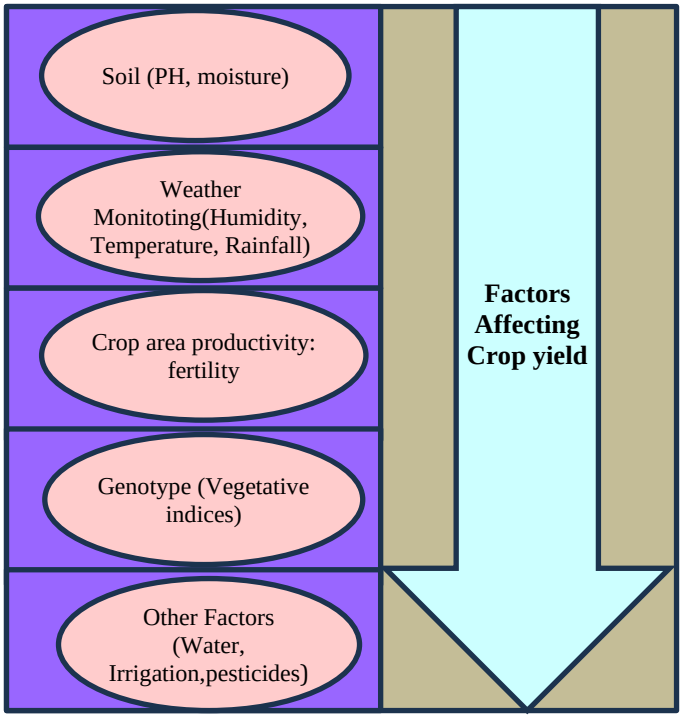
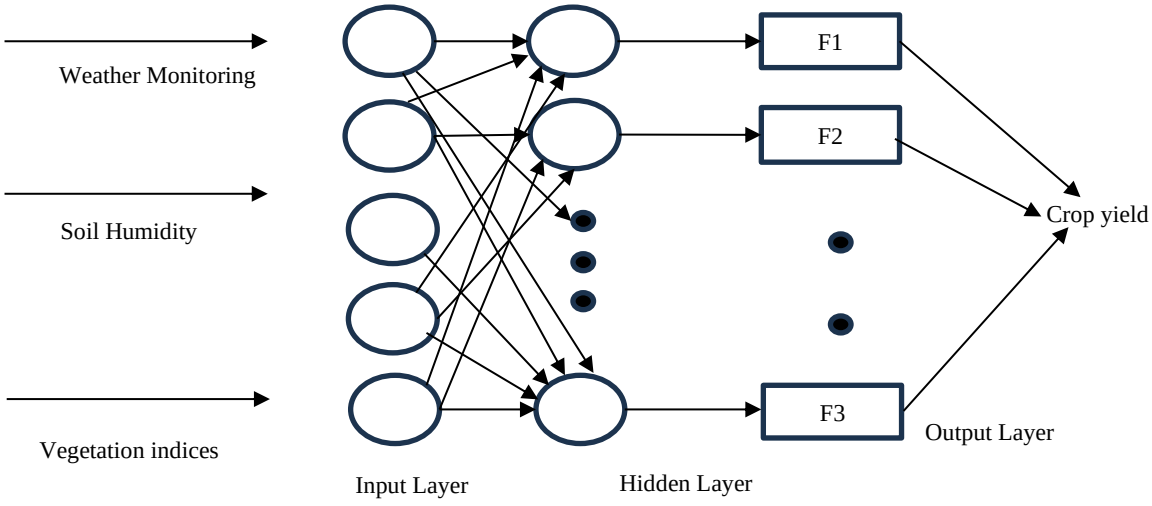
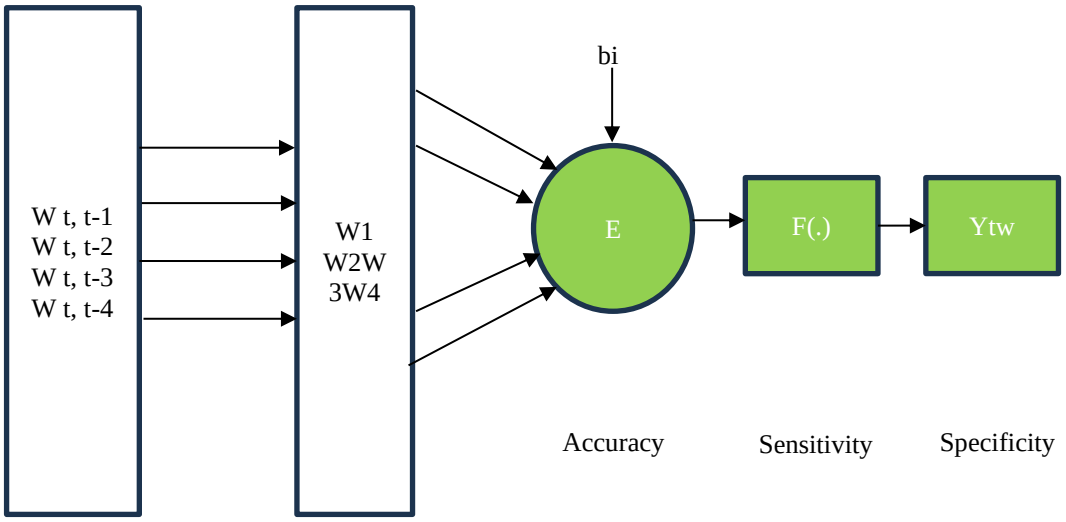


Figure 5: Input parameters for DRL



(a)



(b)

Figure 6(a, b): DQN model

4 Results & Discussion

This research aims to increase yield production by utilizing a variety of machine learning techniques. A dataset related to agriculture is used for experimentation. In order to identify the optimal classification that produces correct projections, classifiers based on machine learning such as Random Forest, XGBoost, Logical Linear Re regression, and neural networks are put into practice. These artificial intelligence methods are implemented utilizing NumPy, Scikit-learn, Pandas, as well as Keras, which are built-in libraries on Python version 3.7 (Jupyter Notebook). Eight gigabytes of RAM and an Intel i-5 CPU are the hardware specifications for implementation (Figures 7-21).

Compared to earlier machine learning (ML) models, the presented ID-RDRL intrusion detection model offers a number of advantages. Listed below are some of the benefits:

(1) The policy, value operation, as well as Q function of the neural networks used for implementing the framework allow the system to quickly and accurately adapt to new networks; (2) The neural network models that are generated are appropriate for dispersed high-performance computer systems (e.g., TensorFlow, Pytorch); (3) the model's level of complexity is decreased as a result of the parameters being much less than those of deep learning networks, as well as (4) the framework is utilized for autonomous learning applications.

The procedure is applied in the manner described below. 1. First, the dataset is preprocessed, involving standardized procedures, cleansing, transformation, and the integration of data. The step of selecting features receives the processed information in order to identify the optimal subset of characteristics.

The dataset utilized in this study's research was first gathered from publicly accessible government websites, https://data.gov.in as well as www.mospi.gov.in. The acquired database includes the following information for the years 2010 through 2018: rainfall, region, area under agriculture, crop names, seasons, manufacturing, as well as yield. Figure 1 shows the crop's output over the previous eighteen years. The agricultural yield for India is predicted by this study using the PCA, Random Forest, as well as Select best Econometric techniques. Additionally, mean absolute mistake, the average squared error, as well as root mean square error are used for validation.

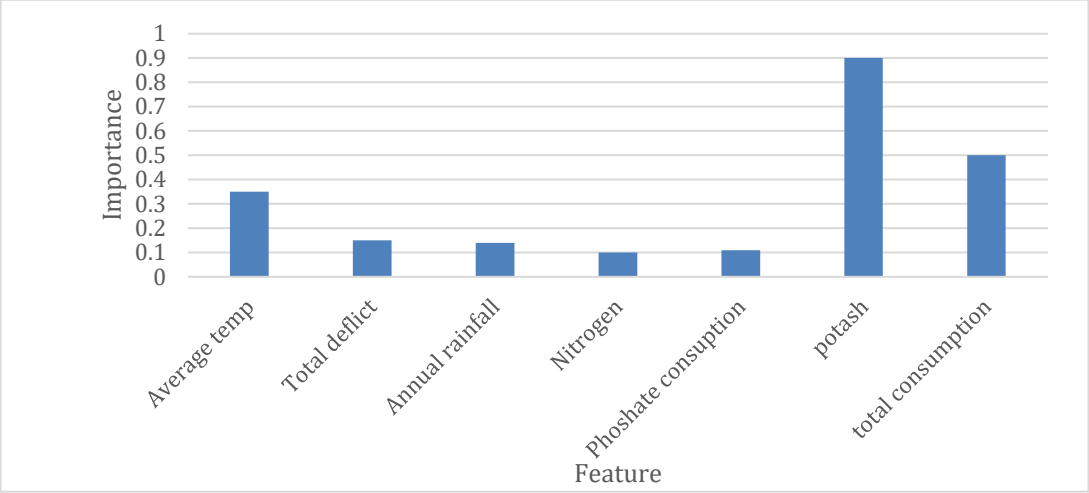


Figure 7: Feature value for rice (PCA)

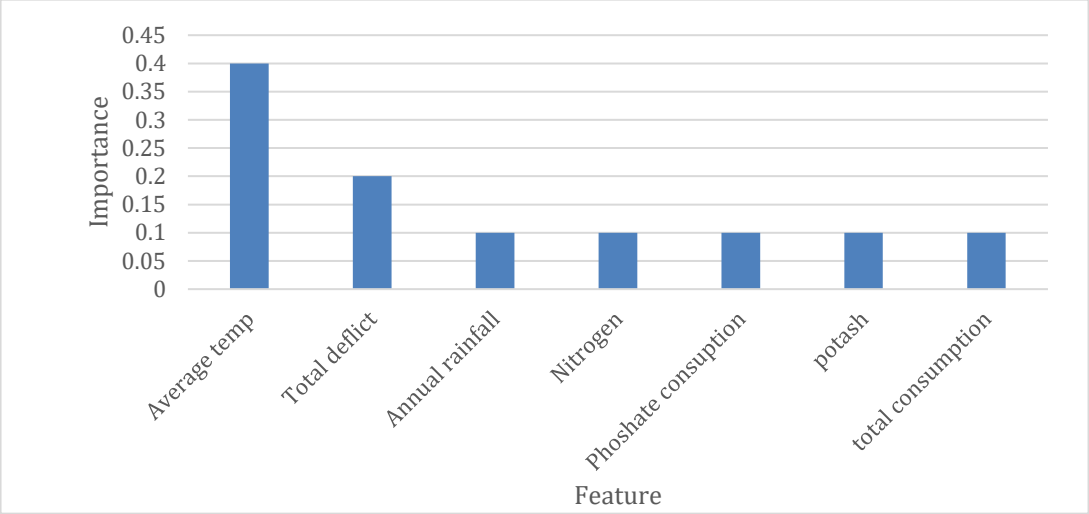


Figure 8: Feature value for rice (Select K Best)

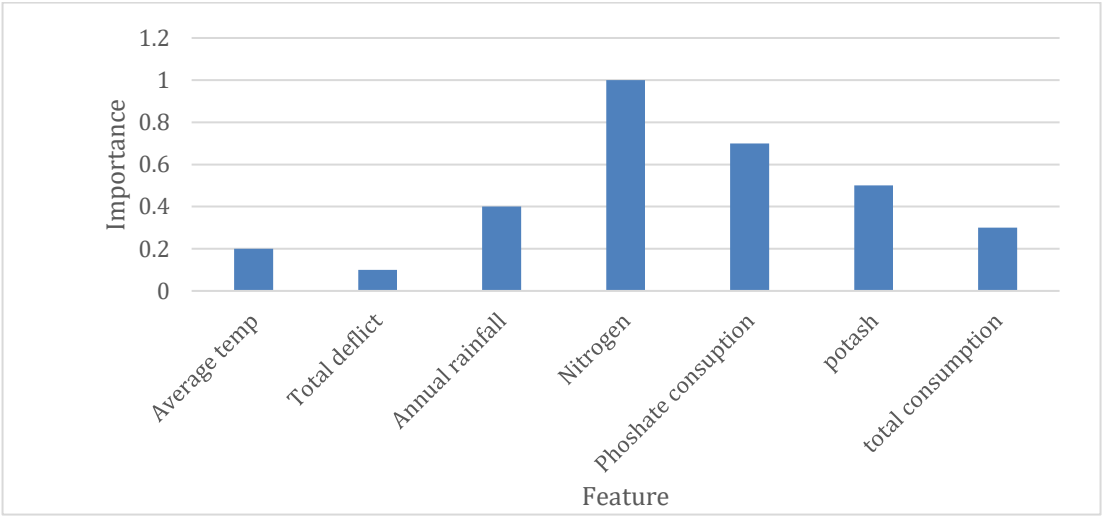


Figure 9: Feature value for rice (random forest)

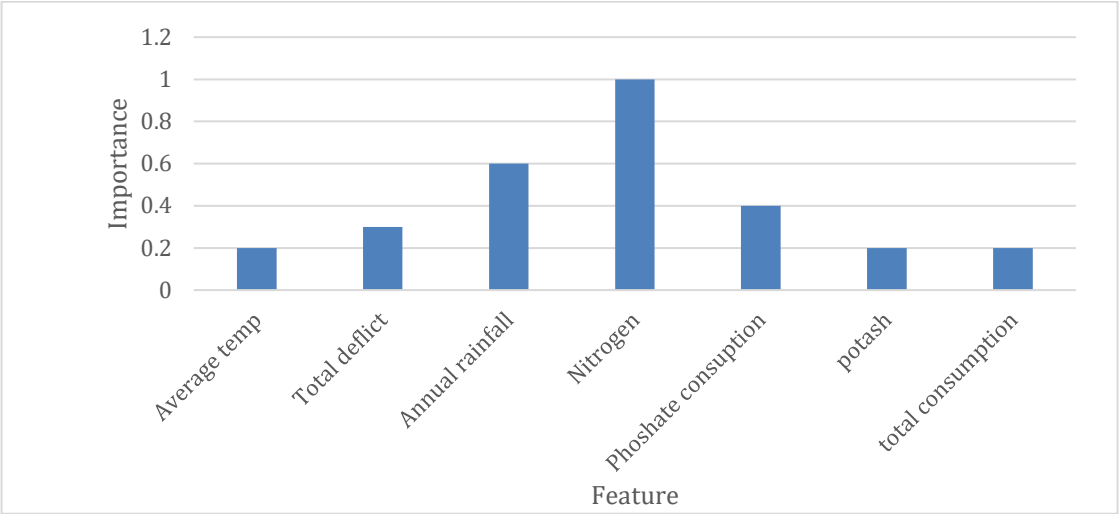


Figure 10: Feature value for maize (PCA)

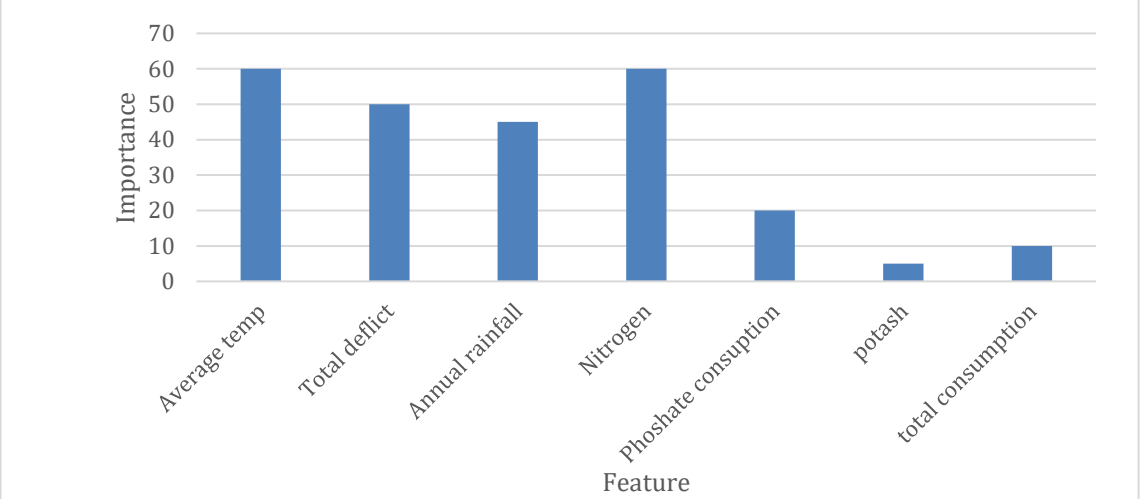


Figure 11: Feature value for maize (select K best)

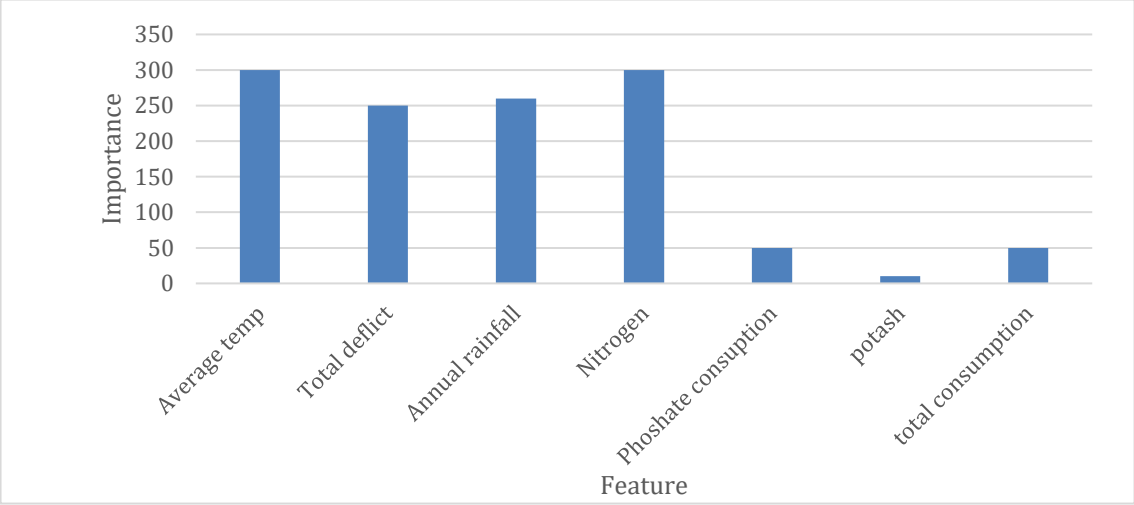


Figure 12: Feature value for maize (random forest)

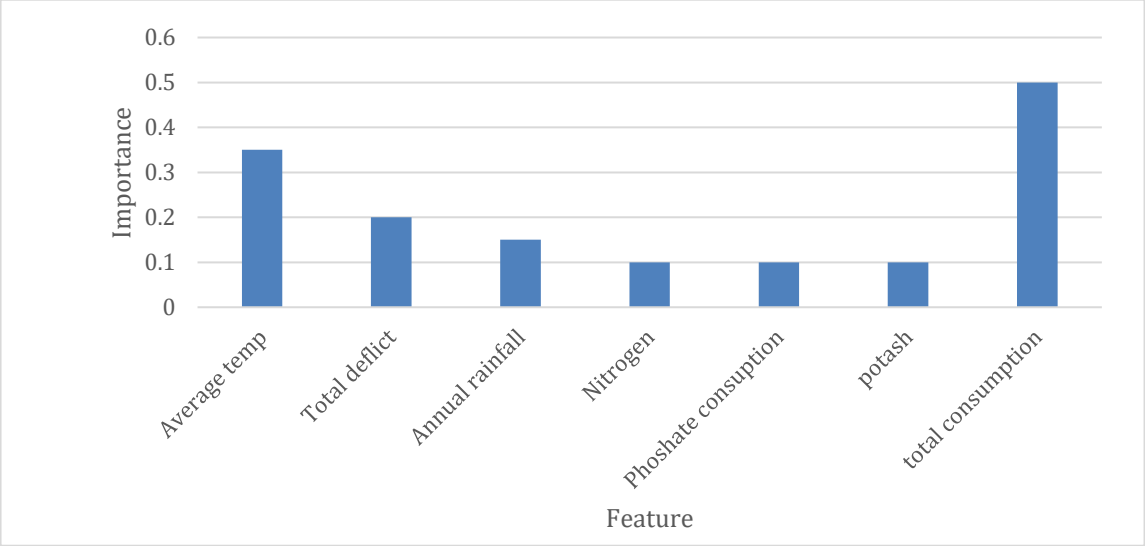


Figure 13: Feature value for wheat (PCA)

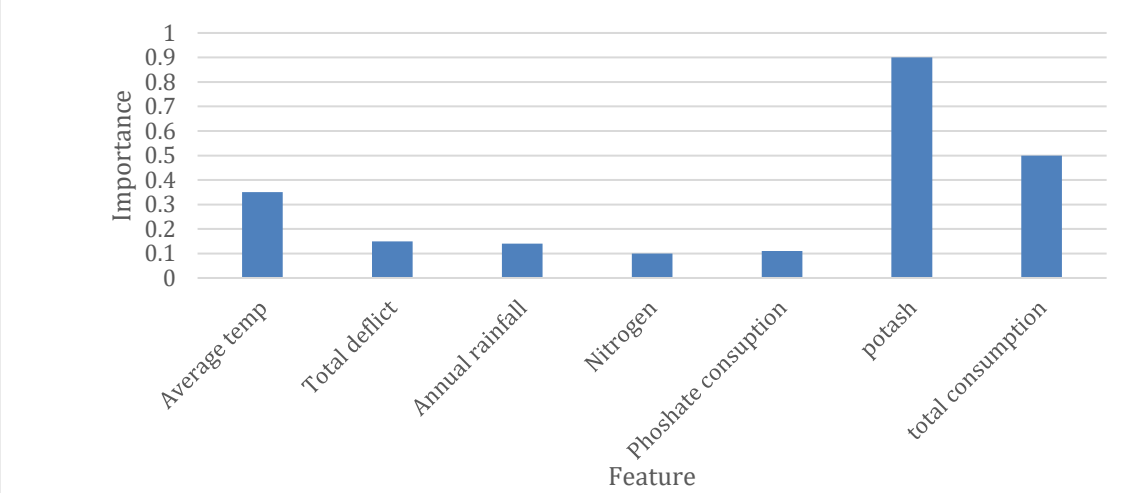


Figure 14: Feature value for wheat (select K best)

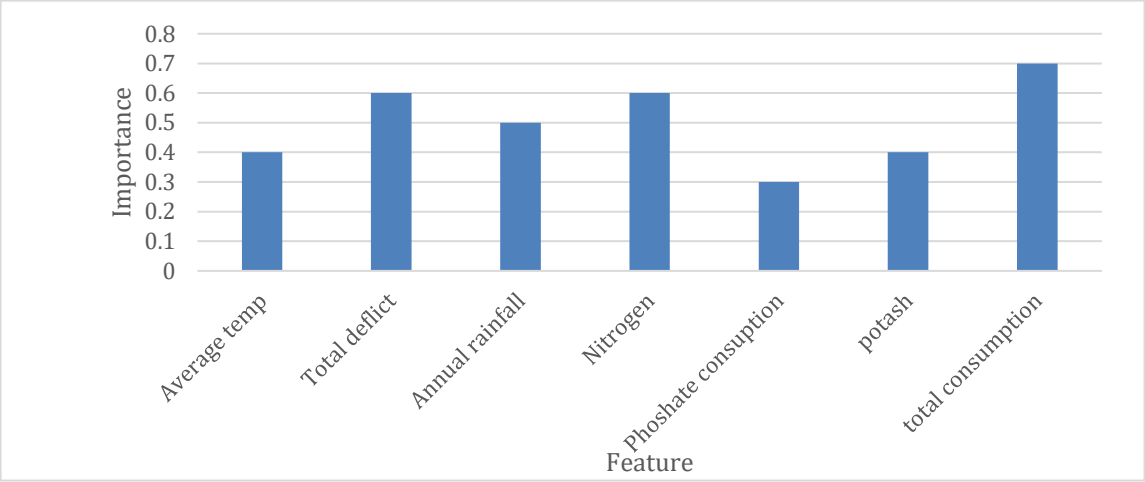


Figure 15: Feature value for wheat (random forest)

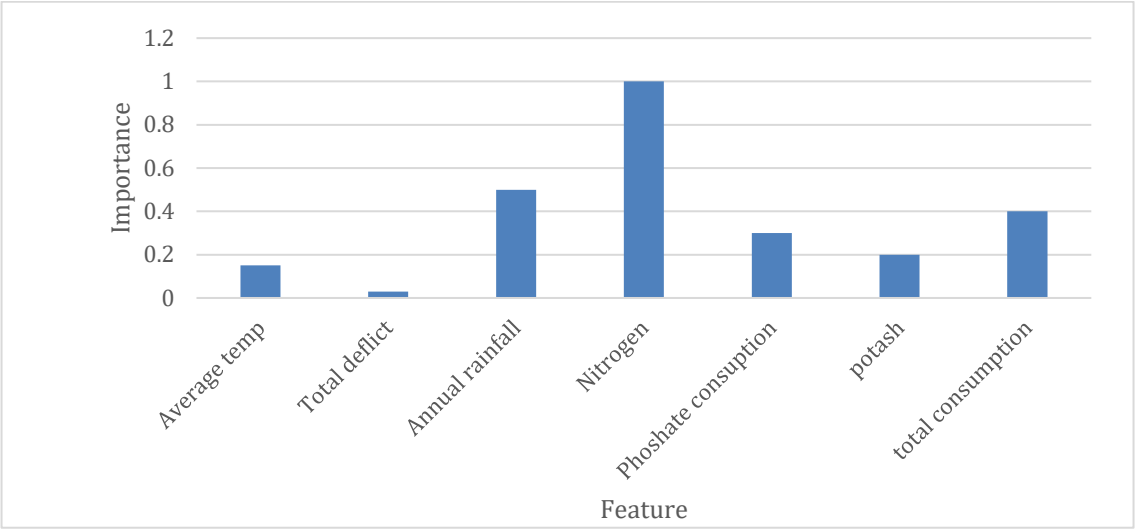


Figure 16: Feature value for cotton (PCA)

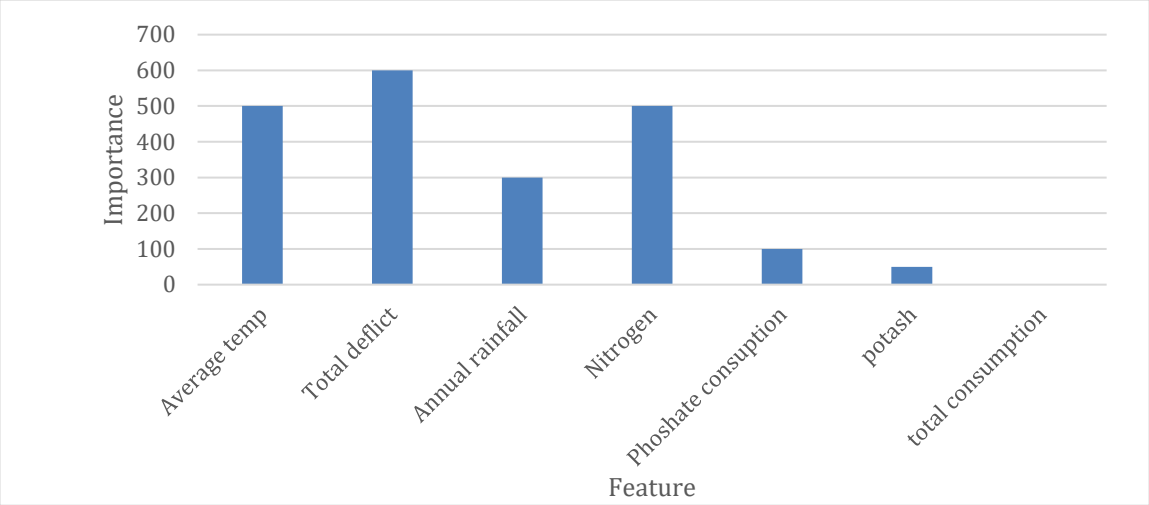


Figure 17: Feature value for cotton (select K best)

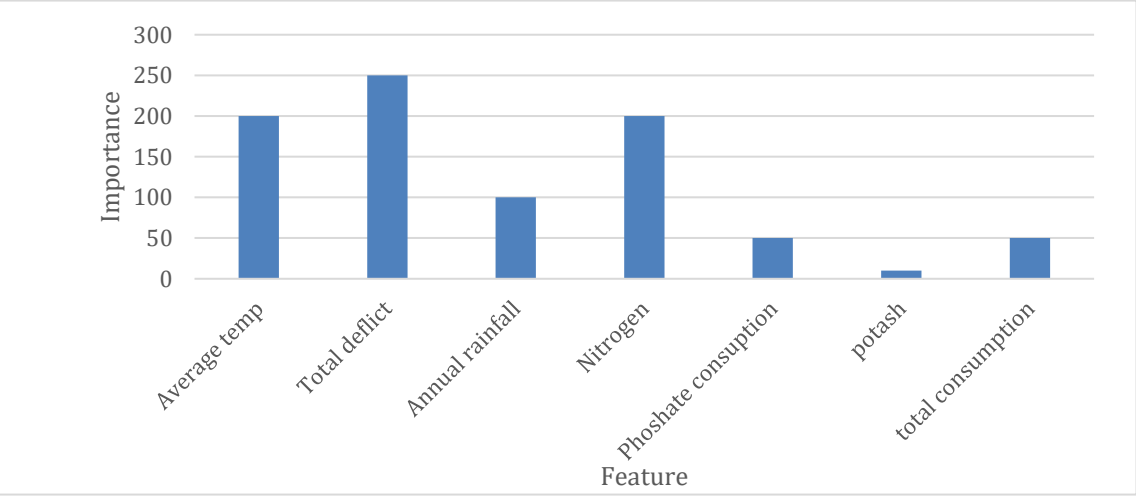


Figure 18: Feature value for cotton (random forest)

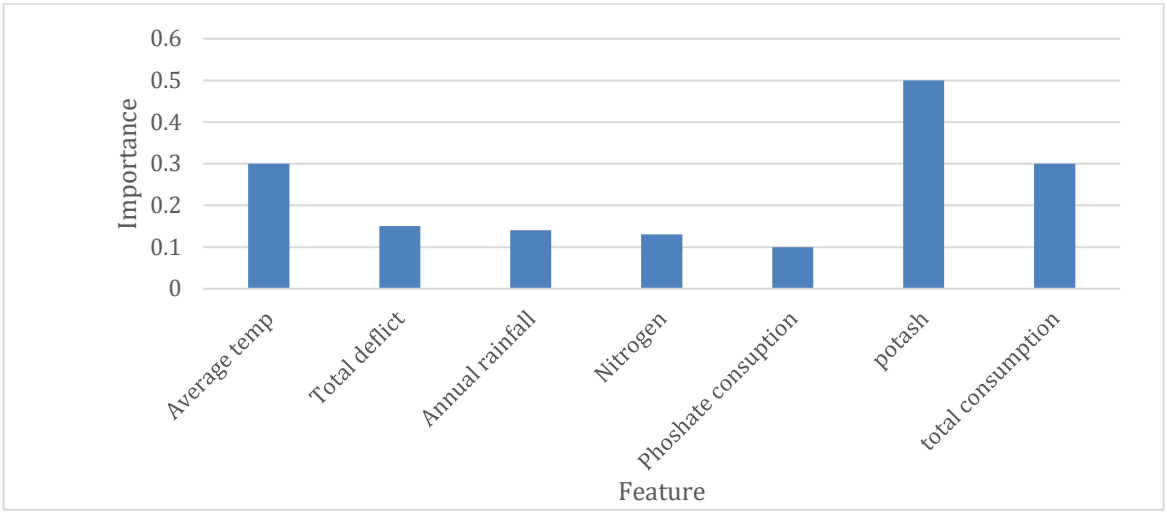


Figure 19: Feature value for soyabean (PCA)

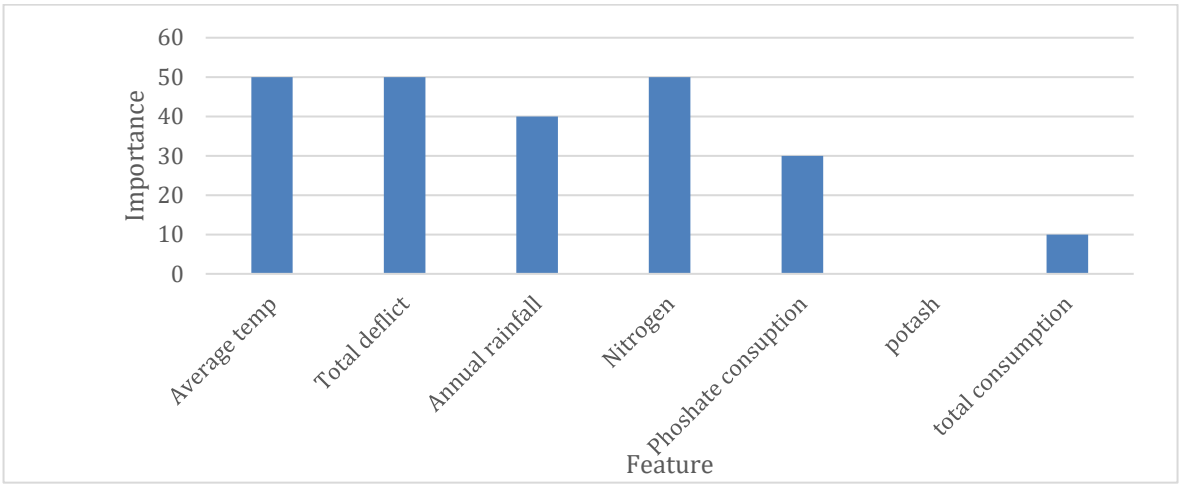


Figure 20: Feature value for soyabean (random forest)

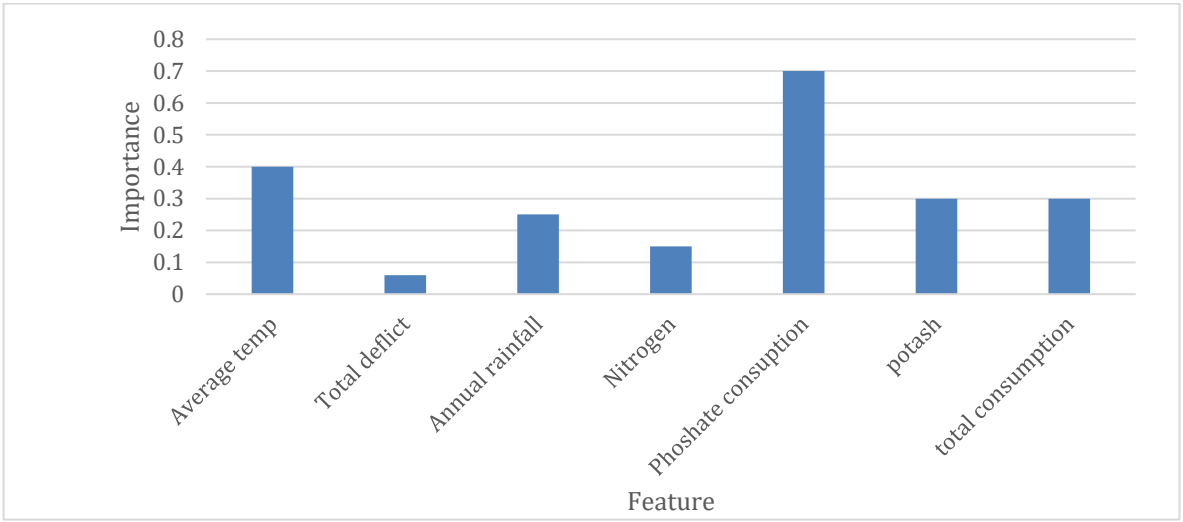


Figure 21: Feature value for soyabean (select K Best)

The Following Three Measures are Used to Assess the Algorithm's Performance

Performance measures are crucial for evaluating the Deep Q-Network's (DQN) learning effectiveness and decision-making precision. They shed light on how well the DQN forecasts the best course of action and optimises cumulative rewards over time. The difference between the predicted and target Q-values is frequently assessed using metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) (Ytw). Better policy selection and sustained convergence result from the network's effective learning to approximate the ideal Q-function, which is indicated by lower error values. As a result, these performance metrics are important markers of the DQN's overall predictive power and training quality.

Mean Absolute Error (MAE)

Using the Mean Absolute Error approach, the difference among two variables that are continuous is determined. When we compare the values of Y and X, which are observational values of identical phenomena, we may determine which states are anticipated and which ones are seen.

The following is the formula for average absolute errors:

Average absolute errors equation (5):

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (5)$$

Mean Squared error

Mean squared is used to get the average of the squared errors between the estimated and actual values.

also known as mean squared deviation.MSE is constantly in the green. Its role is risk.

Mean Squared Error equation (6):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (6)$$

Root Mean Squared Error

RMSE calculates the difference among values that the algorithm predicts and what is observed. The following is the RMSE equation.

Root Mean Squared Error equation (7):

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (7)$$

It is more important to accurately detect crop yield for production. We use F1-score, precision, recall, as well as ROC metrics to measure the model's effectiveness in addition to accuracy, which is one of the metrics to be taken into account.

Confusion matrices, which comprise the following values: TP, TN, FP, and with a grid layout that allows the performance of the model to be seen, are used to generate these metrics. True positives, or TPs, denote accurately expected attack traffic; genuine negatives, or TNs, denote successfully predicted normal traffic; False positives, or FPs, are instances of normal traffic that are anticipated to be attack traffic, whereas false negatives, or FNs, are instances of attack traffic that are projected to be normal traffic. The crucial component is FN; the lower its value, the less likely it is that an intrusion detection system would misinterpret attack communications. This is the goal of our research. Below is a description of the metrics' basic nomenclature as it was previously explained.

Accuracy is the proportion of the overall number of forecasts that the model correctly predicts.

Accuracy equation (8):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Accuracy This statistic measures the percentage of assault traffic that is correctly classified as attack traffic and has the following mathematical definition.

Precision equation (9):

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

F1-Results This statistic is a reconciling average of the model's precision as well as sensitivity, which is an amalgamated version of both metrics. Our work focuses on the metric of improved F1-Scores, which indicates fewer incorrectly classified flows in an imbalanced dataset.

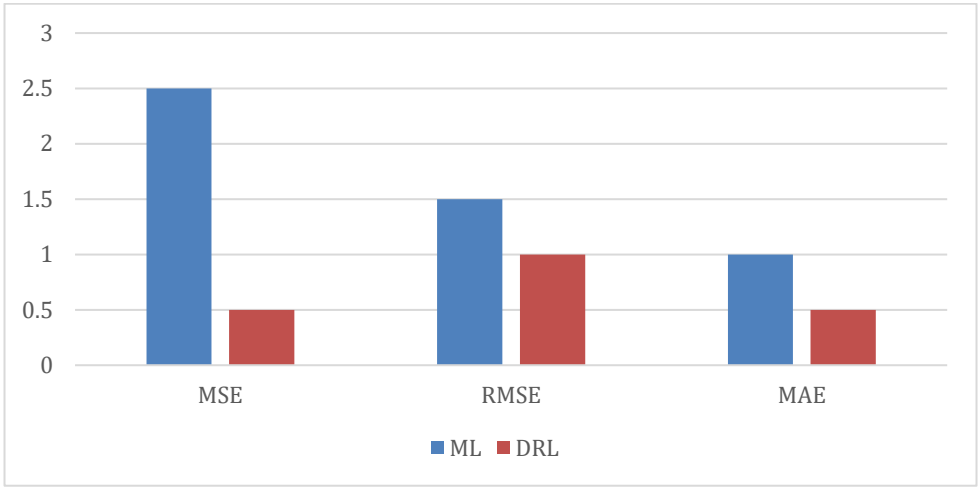
F1-Scores equation (10):

$$F1 = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad (10)$$

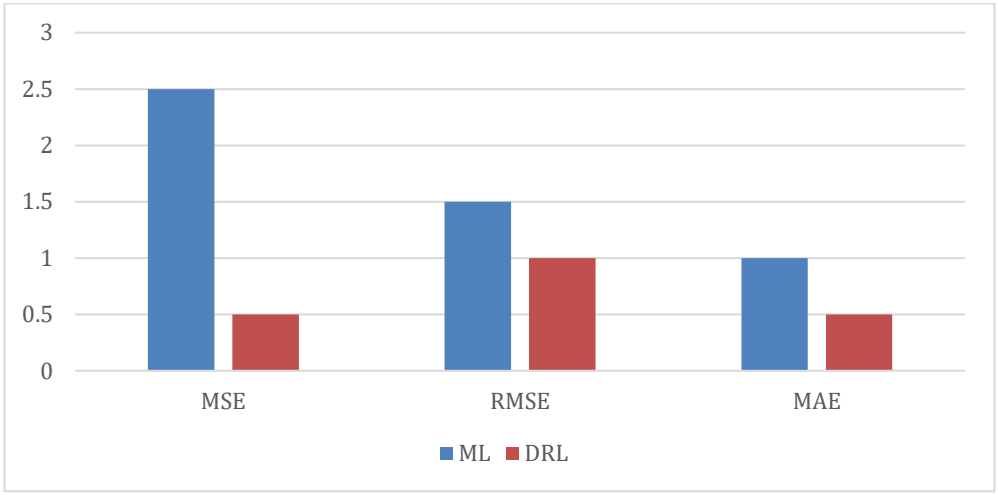
The link between specificity and sensitivity is shown by the receiver's operating characteristics curve (ROC); the better the effectiveness of the model, the wider the region of the curve. The ROC is created using continuous particular parameters and response sensitivities.

Table 1: Algorithms and its description

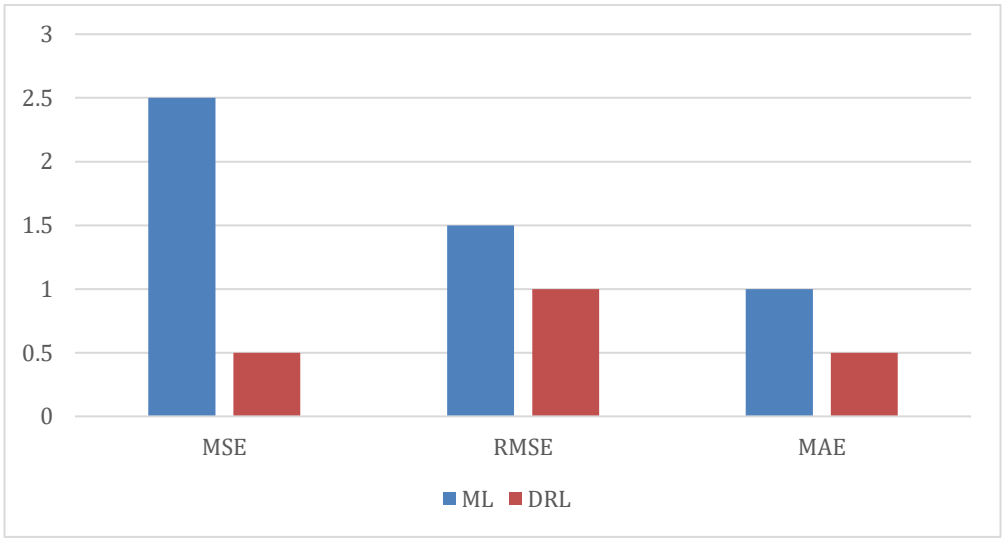
ML Algorithm	**Algorithm Description**
Regression Algorithm	In supervised learning, regression algorithms use training data to establish a connection between input and output, allowing them to predict the numerical value of an unknown input. Polynomial regression, logistic regression, and simple but multiple linear regression are examples of common regression approaches.
kNN	A simple supervised categorization method is kNN. The labeled dataset is initially split into distinct classes in this approach according to their respective outputs. Then, depending on its k-nearest neighbors, a newly collected object is given a class.
RF	Numerous decision tree classifiers are part of the Random Forest ensemble classification system.The majority class projected by many decision tree classifiers determines the last class of a new item.
SVM	SVM is a method for regression as well as classification that creates multi-dimensional borders in the feature space among data points. Using the initial data to split into groups, the SVM's output is anticipated.
RNN	An RNN is a forwarding neural network model in which input layers get feedback from neurons' output layers. Self-loops are another component of the network.
ELM	ELMs are single- or multi-layered feedforward neural networks. Its non-iterative methodology, which allows parameter adjustment to be finished in one pass, makes it suitable for real-time classification and regression issues.
MLP NN	A feedforward artificial neural networks with numerous layers of neurons, MLP NN is inspired by biology. The network's synaptic weights are improved using the training dataset before it is used to generalization.
CNN	The most popular deep neural network is CNN. The system in question is made up of many layers of neurons, and at least a single of the network levels uses the mathematical function convolution rather than matrix multiplication.


3
2.5
2
1.5
1
0.5
0
MSE
RMSE
MAE
ML
DRL

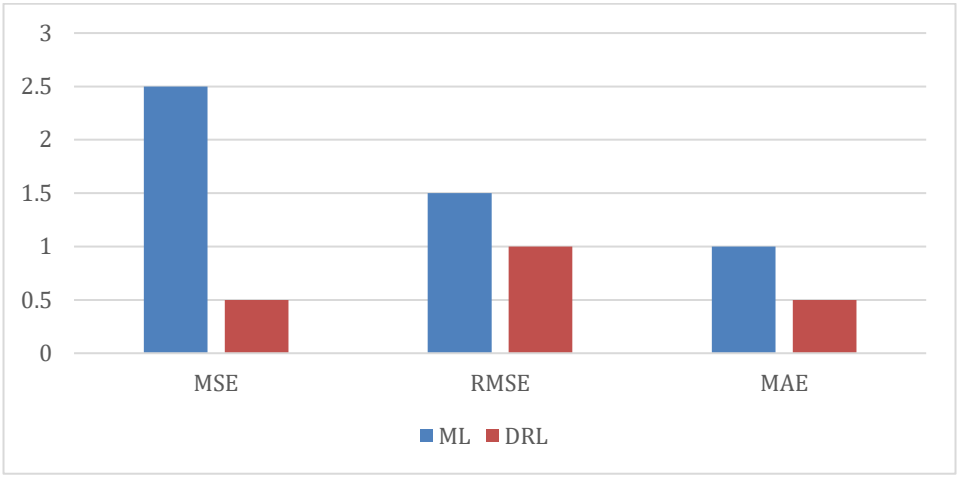
(a)


3
2.5
2
1.5
1
0.5
0
MSE
RMSE
MAE
ML
DRL

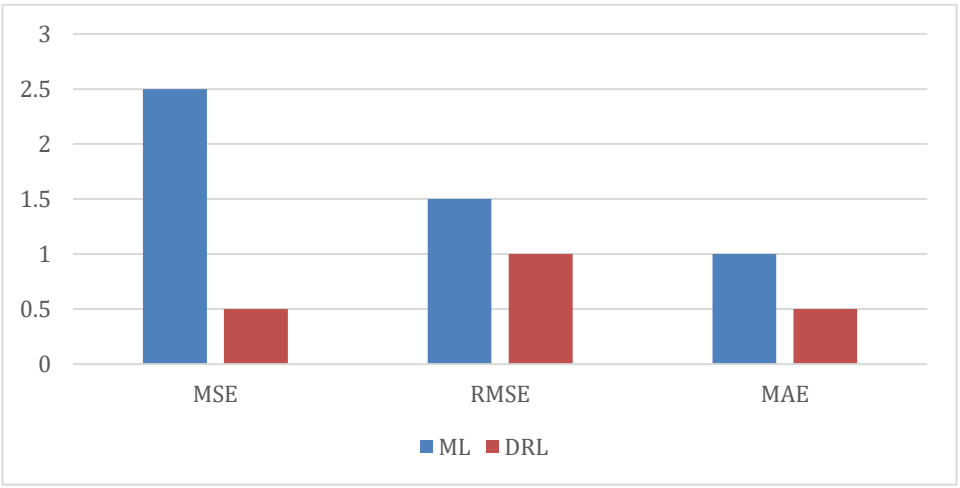
(b)


3
2.5
2
1.5
1
0.5
0
MSE
RMSE
MAE
ML
DRL

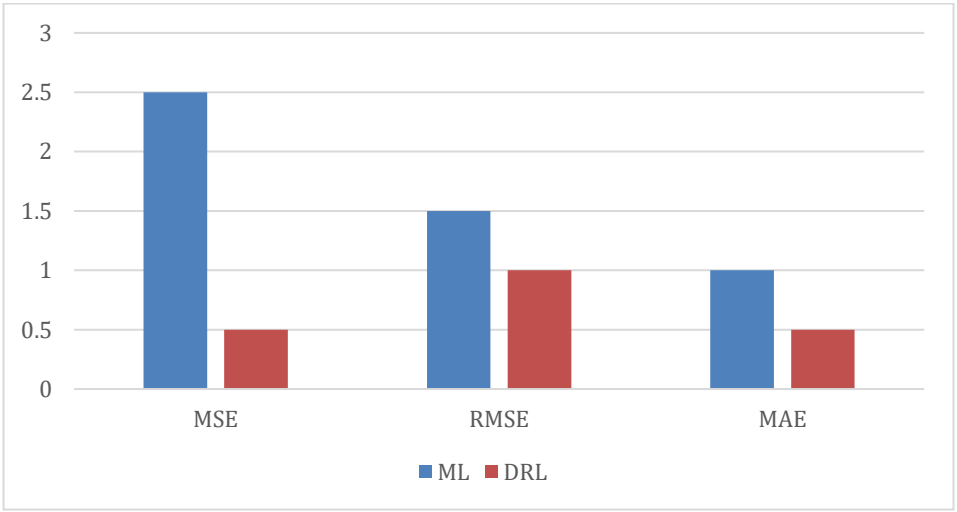
(c)


3
2.5
2
1.5
1
0.5
0
MSE
RMSE
MAE
ML
DRL

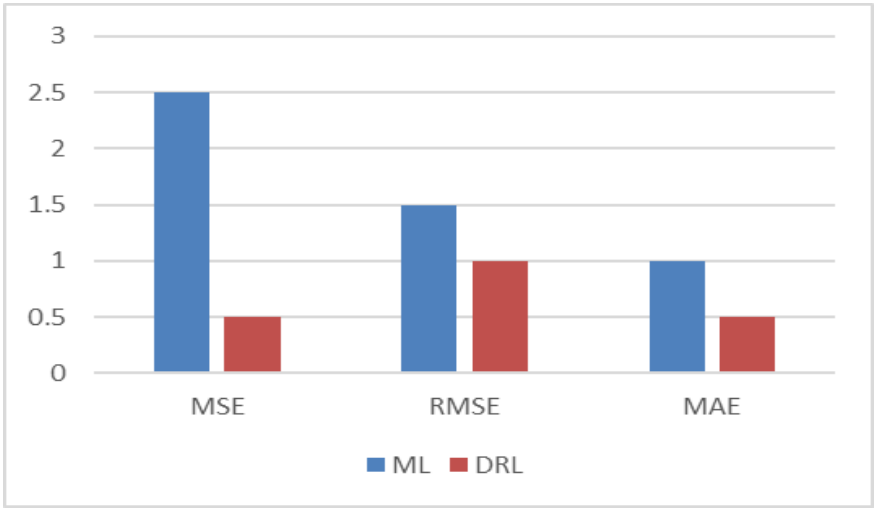
(d)


3
2.5
2
1.5
1
0.5
0
MSE
RMSE
MAE
ML
DRL

(e)


3
2.5
2
1.5
1
0.5
0
MSE
RMSE
MAE
ML
DRL

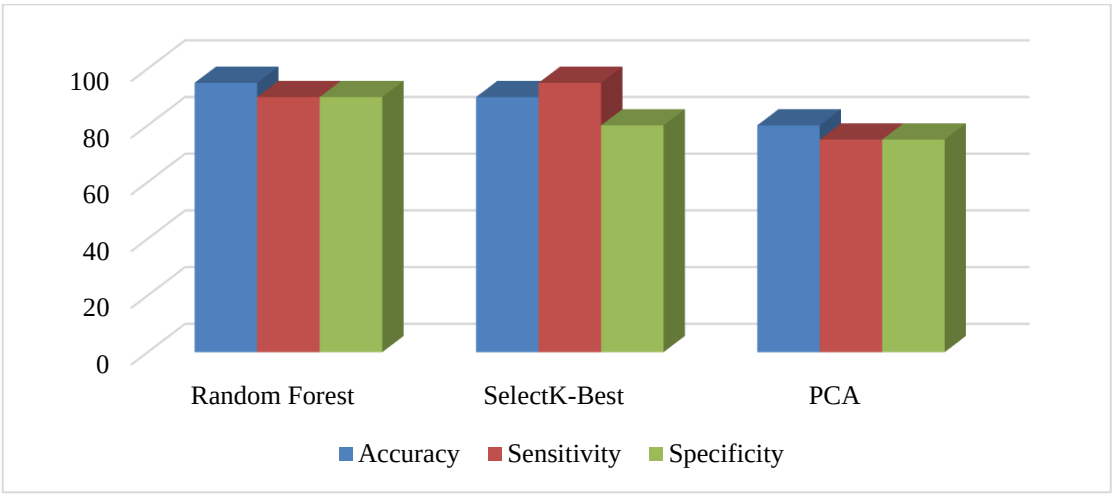
(f)



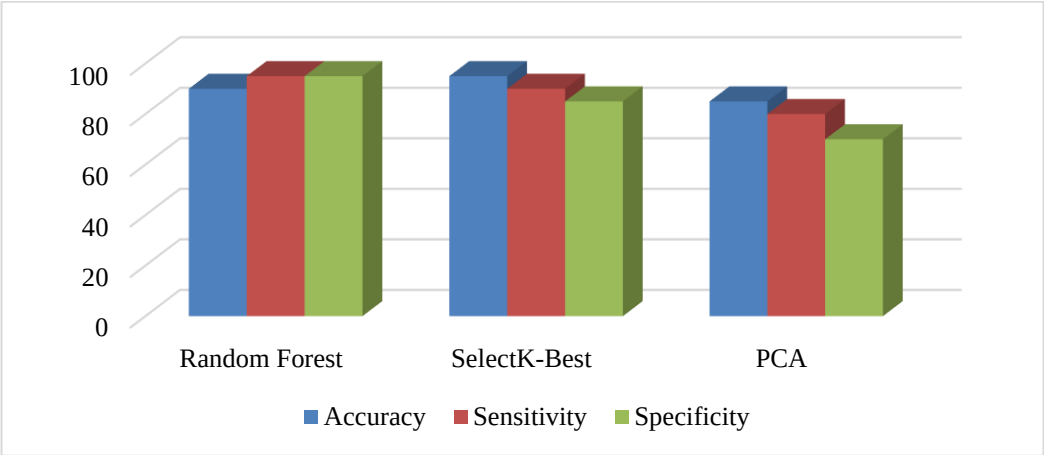
(g)

Figure 22: Error analysis (a). Rice, (b). Wheat, (c). Maize, (d). Cotton, (e). Soyabean, (f). Sugarcane and (g). Corn

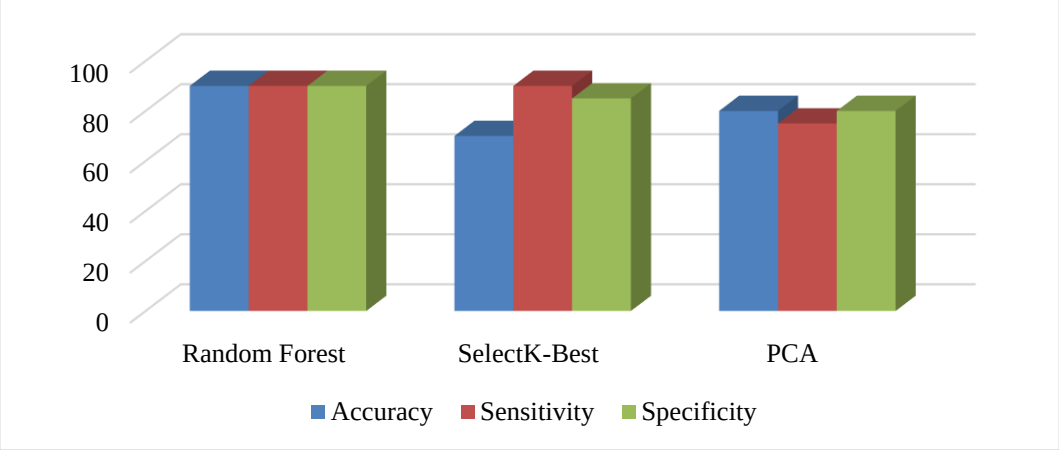
We found remarkable accuracy and precision in the error analysis for wheat yield prediction, the new model exhibited strong predictive stability and reliability. Figure 22 and 23 evaluate the error and accuracy of the model, showing performance for different crops. The Mean Absolute Error (MAE) was measured at 0.5, indicating minimal deviation from actual yield observations, while the Mean Squared Error (MSE) of 0.5 reflected consistent performance with limited variance in predictions. Additionally, the Root Mean Squared Error (RMSE) value of 1.0 demonstrated the model's effectiveness in reducing overall prediction errors. Taken together, these results highlight the model's robustness and dependability, emphasizing its ability to deliver precise and consistent yield estimations when compared to conventional machine learning, model scoring 85% in both, indicating the Deep Q-Network's enhanced capability in accurately identifying true positives and negatives. Moreover, substantial improvements were observed in other key evaluation metrics, including precision, recall, and the F1-score, underscoring the model's overall robustness and predictive effectiveness more accurately than conventional machine learning approaches.



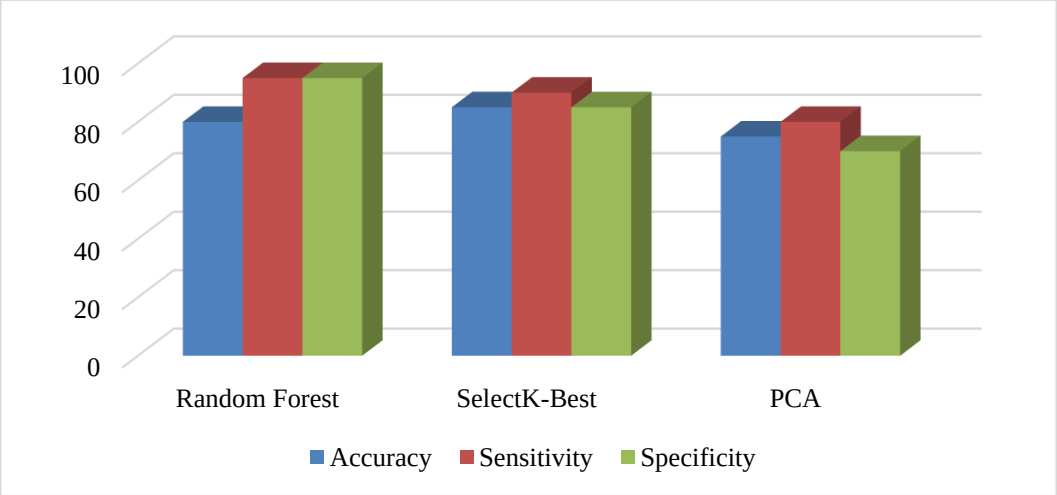
(a)


100
80
60
40
20
0
Random Forest
SelectK-Best
PCA
Accuracy
Sensitivity
Specificity

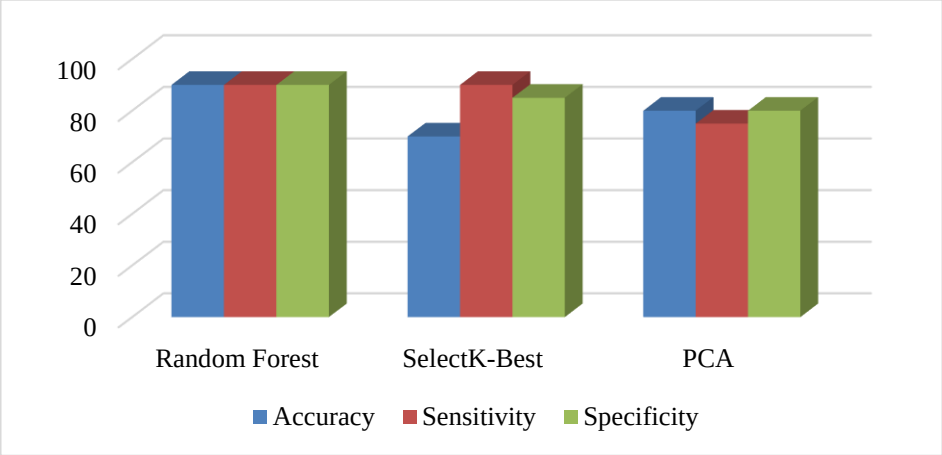
(b)


100
80
60
40
20
0
Random Forest
SelectK-Best
PCA
Accuracy
Sensitivity
Specificity

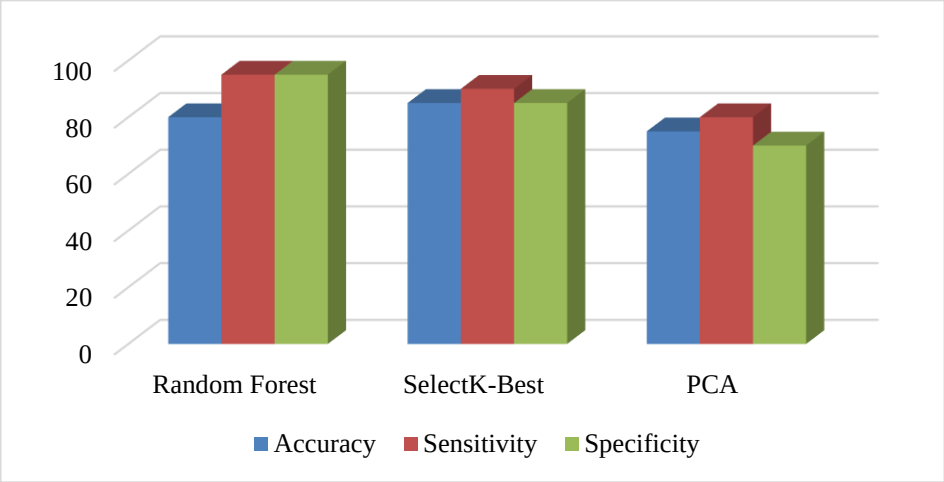
(c)


100
80
60
40
20
0
Random Forest
SelectK-Best
PCA
Accuracy
Sensitivity
Specificity

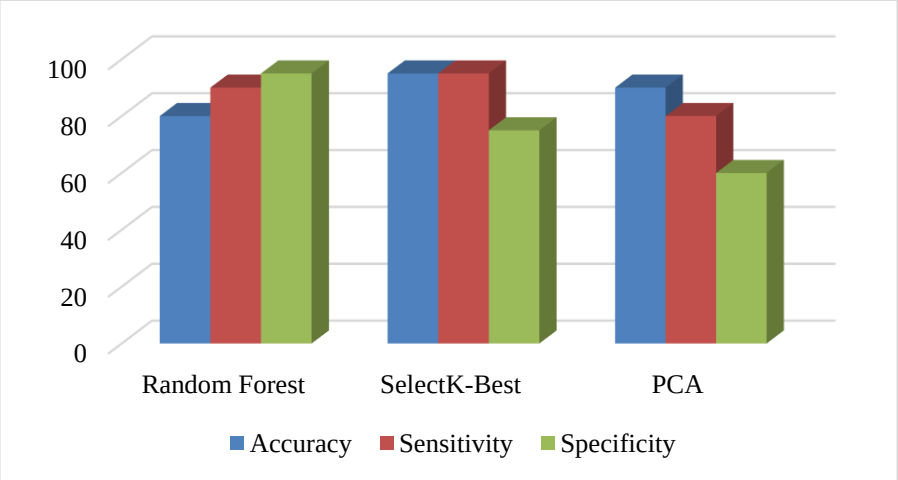
(d)



(e)



(f)



(g)

Figure 23: Accuracy analysis (a). Rice, (b). Wheat, (c). Maize, (d). Cotton, (e). Soyabean, (f). Sugarcane and (g). Corn

In evaluating the performance of the proposed model for yield prediction, the results demonstrated a notable improvement over conventional machine learning approach. The DQN accomplished an accuracy of 92% in predicting rice yield, surpassing the 90% accuracy attained by the Random Forest model. This enhancement was further reflected in the sensitivity and specificity measures, where the DQN consistently outperformed the Random Forest model, indicating its stronger capability to accurately classify yield outcomes and reduce prediction errors is shown in Figure 24.

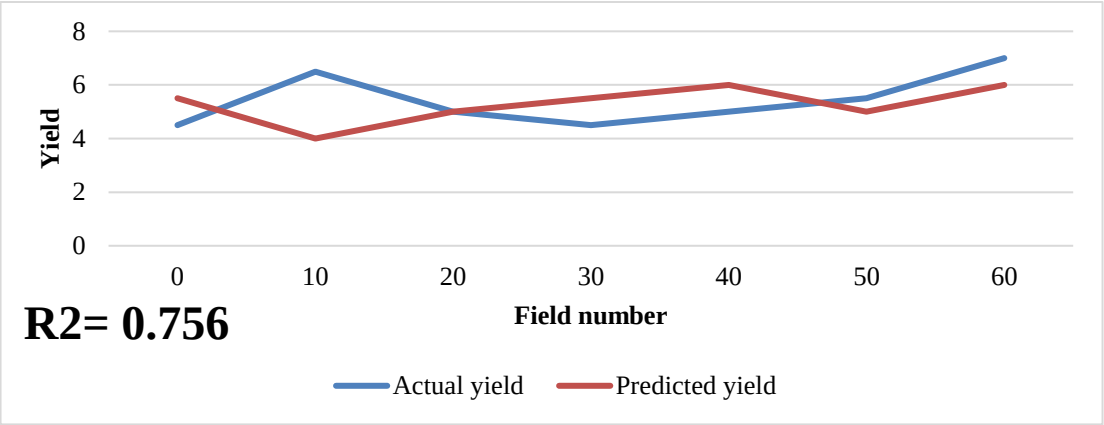


Figure 24: Yield prediction result

5 Conclusion

For countries like India, growing food is the main industry. Nonetheless, the use of technology to agriculture must be prioritized in order to advance precision farming. In this work, we examined the necessity for predicting crop yields as well as the several approaches that were previously proposed in the literature. To anticipate crop productivity, we used deep reinforcement learning algorithms. Accurate agricultural output forecasting can assist authorities as well as organizations with making key policy decisions. We suggested a framework that makes it easier to incorporate different approaches into forecasting crop yields in a flexible way. The integration of ANN with the Grey Wolf Optimizer led to a notable improvement in emission optimization. Quantitatively, our model achieved a 15% reduction in NOx emissions competing with traditional methods. The new method accuracy as well as efficiency were further underscored by its performance metrics. For instance, it outperformed the existing standard models by 10% in terms of accuracy and reduced the error margin to 5%.

Conflict of Interest

The authors declare that no conflicts of interest are associated with the conduct, authorship, or publication of this study.

Ethical Approval

This study was carried out without the need for Institutional Review Board (IRB) approval.

Consent to Participate

All contributors have provided their consent to participate in this research.

Consent for Publication

All contributors have given their consent for the publication of this manuscript.

Data Availability

No vital information, models, or code were created or analysed in this work.

Competing Interests

The authors affirm that they have no competing or conflicting interests in relation to this research work.

Funding

This study was conducted without any external financial assistance or institutional funding.

Author Contributions

All authors made equal contributions to the study's conceptualization, methodology development, data analysis, and manuscript preparation. Each author has thoroughly reviewed the research outcomes and given approval for the final version of the manuscript to be submitted.

Acknowledgements

The authors sincerely thank the Deanship of GITAM University for their unwavering guidance, encouragement, and valuable assistance provided throughout the duration of this study.

References

[1] Abbasi, M., Váz, P., Silva, J., & Martins, P. (2025). Machine learning approaches for predicting maize biomass yield: Leveraging feature engineering and comprehensive data integration. *Sustainability*, *17*(1), 256. https://doi.org/10.3390/su17010256

[2] Alibekova, Z., Ohameed, H., Avezova, U., Muthazhagu, M., Kahorova, K., Udayakumar, R., & Chuponov, S. (2025). Unified GIS-based machine learning method for effective forecasting of disease propagation and resistance in aquatic farming. *International Journal of Aquatic Research and Environmental Studies*, *5*(1), 39-49. https://doi.org/10.70102/IJARES/V5S1/5-S1-05

[3] Anbananthen, K. S. M., Subbiah, S., Chelliah, D., Sivakumar, P., Somasundaram, V., Velshankar, K. H., & Khan, M. (2021). An intelligent decision support system for crop yield prediction using hybrid machine learning algorithms. *F1000Research*, *10*, 1143. https://doi.org/10.1109/ACCESS.2020.2992480

[4] Bali, N., & Singla, A. (2021). Deep learning-based wheat crop yield prediction model in Punjab region of North India. *Applied Artificial Intelligence*, *35*(15), 1304-1328. https://doi.org/10.1080/08839514.2021.1976091

[5] Burdett, H., & Wellen, C. (2022). Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. *Precision agriculture*, *23*(5), 1553-1574.

[6] Elavarasan, D., & Durai Raj Vincent, P. M. (2021). Fuzzy deep learning-based crop yield prediction model for sustainable agronomical frameworks. *Neural Computing and Applications*, *33*(20), 13205-13224.

[7] Elavarasan, D., & Vincent, P. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE access*, *8*, 86886-86901. https://doi.org/10.1109/ACCESS.2020.2992480

[8] Geetha, K. (2024). Integrating Photosynthetic Efficiency Modelling with Crop Yield and Nutritional Forecasting: A Process-Based Approach Toward Global Food Security. *National Journal of Food Security and Nutritional Innovation*, *2*(2), 1-9.

[9] Gong, L., Yu, M., Cutsuridis, V., Kollias, S., & Pearson, S. (2022). A Novel Model Fusion Approach for Greenhouse Crop Yield Prediction. *Horticulturae*, *9*(1), 5. https://doi.org/10.3390/horticulturae9010005

[10] Haghighi, H. F. F., & Far, L. M. (2014). Combining Data Mining and Agricultural Sciences. *International Academic Journal of Science and Engineering*, *1*(2), 1-8.

[11] Hara, P., Piekutowska, M., & Niedbała, G. (2021). Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. *Land*, *10*(6), 609. https://doi.org/10.3390/land10060609

[12] Josephine, B. M., Ramya, K. R., Rao, K. R., Kuchibhotla, S., Kishore, P. V. B., & Rahamathulla, S. (2020). Crop yield prediction using machine learning. *International Journal of Scientific & Technology Research*, *9*(2), 2102-2106.

[13] Liu, Q., Yang, M., Mohammadi, K., Song, D., Bi, J., & Wang, G. (2022). Machine learning crop yield models based on meteorological features and comparison with a process-based model. *Artificial Intelligence for the Earth Systems*, *1*(4), e220002. https://doi.org/10.1175/AIES-D-22-0002.1

[14] Manoj, T., Makkithaya, K., & Narendra, V. G. (2022, February). A federated learning-based crop yield prediction for agricultural production risk management. In *2022 IEEE delhi section conference (DELCON)* (pp. 1-7). IEEE. https://doi.org/10.1109/DELCON54057.2022.9752836

[15] Monica Nandini, G. K. (2024). IOT Use in a Farming Area to Manage Water Conveyance. *Archives for Technical Sciences*, *2*(31), 16–24. https://doi.org/10.70102/afts.2024.1631.016

[16] Moraye, K., Pavate, A., Nikam, S., & Thakkar, S. (2021). Crop yield prediction using random forest algorithm for major cities in Maharashtra State. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN*, 2347-5552. https://doi.org/10.21276/ijircst.2021.9.2.7

[17] Mousa, M. J., & Rasheed, M. K. (2022). Testing the Quality of Forecasting Using the Theils U Statistics for Forecasting the Production Capacity of Crude Oil in Iraq Until 2033. *International Academic Journal of Social Sciences*, *9*(2), 1–7. https://doi.org/10.9756/IAJSS/V9I2/IAJSS0908

[18] Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., & Islam, N. (2022). A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, *14*(9), 1990. https://doi.org/10.3390/rs14091990

[19] Nejad, S. M. M., Abbasi-Moghadam, D., Sharifi, A., Farmonov, N., Amankulova, K., & Lászlź, M. (2022). Multispectral crop yield prediction using 3D-convolutional neural networks and attention convolutional LSTM approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 254-266. https://doi.org/10.1109/JSTARS.2022.3223423

[20] Nyéki, A., & Neményi, M. (2022). Crop yield prediction in precision agriculture. *Agronomy*, *12*(10), 2460. https://doi.org/10.3390/agronomy12102460

[21] Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Hybrid deep learning-based models for crop yield prediction. *Applied artificial intelligence*, *36*(1), 2031822. https://doi.org/10.1080/08839514.2022.2031823

[22] Park, Y., Li, B., & Li, Y. (2023). Crop yield prediction using bayesian spatially varying coefficient models with functional predictors. *Journal of the American Statistical Association*, *118*(541), 70-83. https://doi.org/10.1080/01621459.2022.2123333

[23] Pham, H. T., Awange, J., & Kuhn, M. (2022). Evaluation of three feature dimension reduction techniques for machine learning-based crop yield prediction models. *Sensors*, *22*(17), 6609. https://doi.org/10.3390/s22176609

[24] Pham, H. T., Awange, J., Kuhn, M., Nguyen, B. V., & Bui, L. K. (2022). Enhancing crop yield prediction utilizing machine learning on satellite- based vegetation health indices. *Sensors*, *22*(3), 719. https://doi.org/10.3390/s22030719

[25] Ravi, R., & Baranidharan, B. (2020). Crop yield Prediction using XG Boost algorithm. *International Journal of Recent Technology and Engineering*, *8*(5), 3516-3520. https://doi.org/10.35940/ijrte.D9547.018520

[26] Sadasivam, G. S. (2021). Crop Yield Prediction using Granular SVM. *International journal of recent technology and engineering*, *9*(6), 85-89.

[27] Sajid, S. S., Huber, I., Archontoulis, S., & Hu, G. (2022, June). Integrating crop simulation and machine learning models to improve crop yield prediction. In *2022 17th Annual System of Systems Engineering Conference (SOSE)* (pp. 120-125). IEEE. https://doi.org/10.1109/SOSE55472.2022.9812678

[28] Sajid, S. S., Shahhosseini, M., Huber, I., Hu, G., & Archontoulis, S. V. (2022). County-scale crop yield prediction by integrating crop simulation with machine learning models. *Frontiers in Plant Science*, *13*, 1000224. https://doi.org/10.3389/fpls.2022.1000224

[29] Shafi, U., Mumtaz, R., Anwar, Z., Ajmal, M. M., Khan, M. A., Mahmood, Z., ... & Jhanzab, H. M. (2023). Tackling food insecurity using remote sensing and machine learning-based crop yield prediction. *IEEE Access*, *11*, 108640-108657.

[30] Shetty, S. A., Padmashree, T., Sagar, B. M., & Cauvery, N. K. (2021). Performance analysis on machine learning algorithms with deep learning model for crop yield prediction. In *Data intelligence and cognitive informatics: Proceedings of ICDICI 2020* (pp. 739-750). Singapore: Springer Singapore.

[31] 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. https://doi.org/10.1371/journal.pone.0252402

[32] Sivanandhini, P., & Prakash, J. (2020). Crop yield prediction analysis using feed forward and recurrent neural network. *International Journal of Innovative Science and Research Technology*, *5*(5), 1092-1096.

[33] Sunil, G. L., Nagaveni, V., & Shruthi, U. (2022, July). A review on prediction of crop yield using machine learning techniques. In *2022 IEEE region 10 symposium (TENSYMP)* (pp. 1-5). IEEE. https://doi.org/10.1109/TENSYMP54529.2022.9864482

[34] Umamaheswari, T. S. (2025). Mathematical and Visual Comprehension of Convolutional Neural Network Model for Identifying Crop Diseases. *International Academic Journal of Innovative Research*, *12*(1), 1-7. https://doi.org/10.71086/IAJIR/V12I1/IAJIR1201

Authors Biography

A.K. Saritha is currently pursuing her Ph.D. in Computer Science and Engineering at GITAM University, Bengaluru. Her research focuses on hybrid ensemble learning techniques for crop yield prediction and agricultural data analytics.

Dr.I. Jeena Jacob is a Professor in the Department of Computer Science and Engineering at GITAM University, Bengaluru, and serves as the Ph.D. supervisor of Jeena Jacob with extensive academic and research experience, Dr. Jeena has guided numerous postgraduate and doctoral scholars in developing intelligent systems for data-driven decision-making. Her contributions emphasize the use of advanced computational models for sustainable technological solutions.