"Bio-Agricultural Analysis&Forecasting for Crop Nutrients Management"

Snehal W. Wasankar[1]*, P.M.Jawandhiya[2]

[1]SIPNA College Of Engineering & Technology, Amravati
[2]PLIT, Buldhana
*Corresponding Author

Received: 13.04.2024 Revised : 15.05.2024 Accepted: 23.05.2024

ABSTRACT
The research thrust in preemptive analysis, classification, and validation has been advocating for the development of a framework for making predictive models ever since the need arose for powerful agricultural decision-making tools. Thus, in the current paper, an effort will be made to close the gap among existing avenues within the context of time-series forecasting by incorporating an ensemble, deep learning methodology for classification followed by a detailed analysis of each step entailed in the study, including rigorous validation through ANOVA. This is also one of the limitations, as they show low accuracy and inappropriate validation processes towards dynamic and complex environments, such as an agricultural environment with various scenarios, which require real-time decision support. Instead, the model proposed builds up increased predictive accuracy by initially understanding the critical agricultural parameters such as nitrogen, phosphorus, potassium, temperature, humidity, pH, and rainfall through the use of SARIMAX. This is a preemptive strategy that the input data, in turn, reflects potential future conditions and thus improves the validity for a number of subsequent classifications. This is done for classification with an ensemble of deep learning methods that constitute K-scatter nearest neighbors, random forest, and linear SVC along with logistic regression and multinomial naive Bayes. That is, the classifiers were called upon because these are suitable for high-dimensional, non-linear data, thus giving robust performance for a number of cases. The performance measurements further confirmed using ANOVA (analysis of variance) to test the level of statistical significance of the differences in classifier accuracy. This critical work has greatly bolstered the dependability and accuracy at which decisions in agriculture are made. The sowing recommendation is conditioned in this way before the predicted values, increasing the accuracy of predictions and supplying, through validation, a comprehensive framework as a basis for future studies in predictive modeling and classification in agriculture sets.

Keywords: Preemptive Analysis, SARIMAX, Ensemble Learning, Agricultural Forecasting, ANOVA Validation, Scenarios

1. INTRODUCTION

Predictive analytics and machine learning have become vital components in this fast-paced evolution taking place within agricultural technology to attain the best crop production and resource management. The increased rate of the world's population directly puts higher demands on food production, which shall hence require adaptation of advanced techniques to raise agricultural efficiency and sustainability. The traditional methods of crop management apply little use of historical data and reactive measures; these are poorly equipped against the new complexities of agricultural ecosystems. Predictive models that forecast future conditions will be important in overcoming these limitations, together with robust classification frameworks. The approach with which this paper is involved in making decisions in agriculture combines preemptive analysis using the SARIMAX model, which is an acronym for Seasonal Autoregressive Integrated Moving Average with Exogenous variables, with ensemble deep-learning methods for classification, and rigorous validation by ANOVA. The application of SARIMAX is due to the fact that it has the capability of handling time-series data with seasonal components, which is very critical in nature for agriculture, where environmental conditions usually vary seasonally. This means that the model will be based on the most accurate and relevant data while dealing with classification if it can project the critical parameters like soil nutrients, temperature, humidity, pH level, and rainfall.

The classification phase uses an ensemble consisting of K-Nearest Neighbors, Random Forest, Linear Support Vector Classifier, Logistic Regression, and Multinomial Naive Bayes. The adopted models were used en bloc so as to exploit their strengths, which further complement one another very well in high-

dimensional and nonlinear data typical of agricultural data sets. KNN is known for its simplicity and effectiveness in smaller datasets, while Random Forest provides robustness against overfitting by averaging multiple decision trees. The choice of Linear SVC and Logistic Regression was based on the interpretability and computational efficiency of these two algorithms; both could be deployed in real-time applications. Multinomial Naive Bayes, even though primarily used for text classification, was included because it can handle categorical data, which is often part of agricultural datasets and samples. ANOVA is one of the statistical methods which validate the proposed model, pointing out whether there exist statistically significant differences between the means of multiple groups. In the present paper, it is used to find out the performance differences of different classifier results. This is a stringent step of validation, very critical in ensuring the reliability and robustness of the model under agriculture scenarios in the real world with serious prediction errors.

One of the main challenges in agricultural prediction and classification is the dynamic nature of the environment. This change is affected by several factors, such as weather patterns, soil conditions, and human factors. Such complexities go mostly unrepresented in the older modeling techniques, thereby rendering less-than-optimal decisions that may negatively affect crop yield and resource use. In light of this, the proposed method incorporates SARIMAX to make pre-analyzed forecasts of seasonal variations and exogenous factors, hence giving a truer picture of the future conditions. Besides, an ensemble approach for classification makes the model resilient on different agriculture data sets. Ensemble methods tend to improve predictive accuracy and lower variance of the model. Such a model combines the strengths of the different classifiers, whereby it is hence stronger than any one algorithm: where one algorithm might be weak, another one could complement it. This aspect will become much more significant in agriculture, in which data is noisy, incomplete, and misclassification could have bad consequences. A validation process through ANOVA is an added rigor to this model. While high accuracy is mostly the objective in many machine learning studies, not much emphasis is put on the statistical significance of the results. However, when one applies this in real life, especially in critical areas like agriculture, one must be sure that the performance differences observed are actually not because of random chance. The approach takes the help of ANOVA to make the model statistically sound. This enhances and boosts the confidence in the results of different classifier performances.

The methodology proposed brings in an appreciable overture to the already existing approaches of agricultural forecasting and classification. Most traditional approaches to the tasks in question are based on some historical data and some unduly simplistic predictive models that can't capture the modern complexities of the agricultural systems. By integrating elements of preemptive analysis, ensemble learning, and rigorous validation, the model proposed gives a more accurate, reliable, and robust solution for agricultural decision-making. This may usher in far better practices in crop management that lead not only to higher yield but also to better resource utilization, thus finally paving the way for more sustainable agriculture estates. In other words, SARIMAX contributes to predictive analysis, ensemble learning contributes to classification, and ANOVA validation; all these put together give a comprehensive framework for agricultural forecasting. It improves the deficiencies of the existing approaches and provides a more accurate and reliable solution for optimizing crop production and resource management. The model proposed here is not only sound from a technical point of view but full of practice, likely to contribute much to the agricultural industry. Keeping in view the ever-growing demand for the production of food, advanced techniques as discussed in the paper would have to be put in place to ensure that food security was maintained.

Motivation & Contribution

This work is motivated by the inadequacy of most existing agricultural models, which deal with forecasting and classification in modern agricultural systems, because of their inability to incorporate the dynamic and complex nature. Traditional approaches are often reactive, thus based on previous data, hence unreliable in a well-seasonal and extra driven environments like climate change. Agriculture has become over-reliant on accurate predictions regarding crop management and, hence, requires advanced techniques which can predict the future conditions and provide reliable recommendations to farmers and stakeholders. It is further driven by the urge to come up with a method that would not only increase predictive accuracy but would also be backed by a high degree of validation process to guarantee the reliability of the results, hence filling a missing link in the current state of agricultural technology.

The contribution of the research is manyfold. First, this research proposes a new integration using SARIMAX for analysis on a preemptive basis that makes this model forecast critical parameters of agriculture with high precision. This is very important in agriculture where most environmental factors, like temperature, humidity, and soil nutrients, keep on varying at different instances; therefore, forecasting accurately is very important to optimize crop yields. This study considers an approach that

incorporates SARIMAX so that the input data used for classification takes into account the most plausible future state, hence improving the accuracy of the classification results. The agricultural data samples were classified with the service of an ensemble of deep learning techniques: KNN, Random Forest, Linear SVC, Logistic Regression, and Multinomial Naive Bayes, in this research. Itself, the ensemble approach exploits the strengths of each classifier to offer a robust and reliable classification framework able to capture all diversity and complexity in agricultural datasets. Finally, this research insists on validation, using ANOVA for checking the statistical significance of the differences in classifier performance. This stringent validation step ensures that the predictions from the model are not only accurate but also statistically sound, hence providing a firm basis for making decisions within any agricultural application. The integration of such techniques hence marks a milestone in this field by giving a holistic solution to the failures of existing models working on more sustainable and efficient agricultural practices.

2. Review of Existing Models for Yield Analysis

Agritronics, itself a relatively new space, has begun to see quite some development toward yield prediction and yield monitoring, supported by the integration of machine learning, remote sensing, and advanced data processing techniques. Growing global concerns related to food security and sustainable agriculture have put a big impetus on pursuing more accurate and reliable methodologies toward the prediction of crop yield and judgement of crop health, as well as optimization of agricultural practices. This is clearly reflected in the diverse array of methodologies showcased within the reviewed body of literature, which were oriented toward addressing the complex challenges inherent in agricultural data analysis. Among the very different technological approaches applied, one finds traditional machine learning taken to the extreme, with approaches such as Support Vector Machines and Random Forests, as well as more advanced approaches, including deep learning, with Long Short-Term Memory networks, Convolutional Neural Networks, and more. Additionally, such methods are strengthened by data fusion techniques and sophisticated statistical tools, such as the Bayesian posterior-based Ensemble Kalman Filters, the cross-wavelet transform, increasing predictive capability, being adaptable in different agricultural environments. A first key unifying theme that emerges from this review centers around the integration of machine learning algorithms and remote sensing data. Remote sensing, through these platforms with the help of sensors available on MODIS, Sentinel, and UAVs, has produced near-high-resolution data in relation to crop condition, soil moisture, and other environmental conditions that play a role in yield prediction with reasonable accuracy. For instance, in the study by Shafi et al. [1], integration of remote sensing data with machine learning models has been shown to greatly improve prediction of wheat yield, especially in countries where food insecurity is a serious problem. For example, Ji et al. [8] integrated data from Sentinel-2 into the CASA-WOFOST model for high accuracy at the field scale in crop yield estimation. These present studies simply highlight the increasing reliance on remote-sensed technologies giving near real-time and high-quality data such that it is integrated with machine-learning models to yield granular insights in agriculture. Limitations on their effectiveness generally arise from the quality and resolution of the remote-sensing data involved, specifically being pointed out by certain studies where the model's capability in capturing fine-grained changes in crop condition was restricted by the lower spatial resolution.

Another important line of research is in the application of deep learning models that have been tailored for time series data or complex feature relationships. For instance, in interpretable LSTM networks for crop yield, Mateo-Sanchis et al. [3] proved significant contributions from the model's variables to the final choices using SHAP values. This strategy not only enhanced the prediction accuracy but also mitigated one of the most criticized points for deep learning models, including the black-box nature, by making the results more interpretable to the end users-mostly the farmers and decision makers in the agricultural sector. In the same way, Zhang et al. applied deep learning to improve yield estimation from multisource temporal drone imagery by the use of an attention mechanism and transfer learning strategies. These approaches emphasize the strength of deep learning in handling agricultural data, which are generally high-dimensional and non-linear; nevertheless, they indicate several issues related to the high requirement of training data and computation resources, hence making them less applicable in resource-constrained environments. Bringing knowledge from domain experts into the setting by embedding crop growth models and models of other biological systems also becomes a general tenet for quite a number of works. For example, Huang et al. illustrated that assimilating GLASS LAI information in a crop growth model using an Ensemble Kalman Filter improved the accuracy of winter wheat yield estimation by capturing the biological dynamics underlying crop growth. This model integrates machine learning techniques with agronomic expertise, ensuring that the predictions are based on the true biological aspects of crop development, which increases the reliability of the results. However, most of the models need major calibration and validation, as in Nikaein et al. [22], where the DSSAT model was expected to

be very carefully calibrated so that it could go along with SAR data in estimating crop growth. These results indicate that although including crop growth models can be effective in significantly enhancing the predictive accuracy, such models add an extra layer of complexity that needs to be managed and controlled. Besides predictive modeling, there have been various studies that have concentrated on the optimization of agricultural practices through better, more precise recommendations based on multisensor data fusion. Reyana et al. [10] have elaborated how the amalgamation of multiple sources of sensor data in machine learning models can lead to faster crop yields and provide actionable cultivation advice for the farmer. Munaganuri and Rao [20] further proposed the PAMICRM model through multimodal image analysis in accurate irrigation management to increase sustainability in agricultural practices. These studies demonstrate the potential for multisenor data fusion to advance decision making in agriculture but also outline the associated challenges related to data synchronization and integration across heterogeneous data sources.

Table 1. Empirical Review of Existing Methods

Reference	Method Used	Findings	Results	Limitations
[1]	SVM, Regression	Integrated remote sensing and ML for crop yield prediction.	Achieved significant improvements in wheat yield predictions.	Limited by the quality of UAV and VI data samples.
[2]	Bayesian Posterior-Based EnKF	Assimilated GLASS LAI into a crop growth model.	Enhanced winter wheat yield estimation accuracy.	Requires high computational resources.
[3]	LSTM, SHAP Values	Applied interpretable LSTM for crop yield estimation.	Provided both accurate predictions and model interpretability.	Limited generalizability to different crops.
[4]	Gradient Boosting, Logistic Regression	Integrated meteorological and pesticide data samples.	Improved accuracy in forecasting crop yields.	Pesticide data variability affects predictions.
[5]	MODIS, Data Fusion	Assimilated Earth observation data for smallholder agriculture.	Enhanced yield estimation in smallholder systems.	Dependent on MODIS spatial resolution limitations.
[6]	Random Forests, GEE	Mapped complex crop rotation systems.	Accurate crop rotation mapping considering intensity and diversity.	Limited by Sentinel-1/2 data availability.
[7]	SVM, RF, CNN	Fused climate and NDVI data for wheat yield prediction.	Achieved high accuracy in wheat yield predictions using fused data samples.	Relies on the availability of high-quality NDVI data samples.
[8]	CASA-WOFOST, NDVI	Assimilated Sentinel-2 data into crop yield models.	Achieved precise yield estimation at field scales.	High temporal resolution data required for best results.
[9]	LSTM, Time Series	Developed TAYP model for improving crop productivity.	Improved time series analysis for yield prediction.	Requires extensive training data for effectiveness.
[10]	Multisensor Data Fusion, ML	Applied multisensor data fusion for crop yield acceleration.	Enhanced crop yield predictions with integrated sensor data samples.	Potential sensor data synchronization issues.
[11]	CROP-DualGAN, Hyperspectral	Improved LAI estimation using GANs.	Achieved high accuracy in LAI estimation.	Model complexity may limit scalability.
[12]	Decision Trees,	Combined regression	Provided accurate	Decision trees may

	Deep Learning	and deep learning for yield prediction.	yield predictions across various crops.	overfit in noisy data samples.
[13]	Hyperspectral Imaging, SVM	Classified crop types using hyperspectral imagery.	Improved crop type classification accuracy.	SVM limitations in handling large datasets.
[14]	Cross-Wavelet Transform, VTCI	Enhanced wheat yield estimation using VTCI and LAI data samples.	Achieved precise wheat yield estimation through enhanced feature extraction.	Requires high-quality wavelet transform data samples.
[15]	DL, Transfer Learning	Compared attention-based DL and transfer learning for yield estimation.	Demonstrated superior yield prediction using multisource temporal imagery.	Transfer learning models require extensive tuning.
[16]	Deep Learning, MODIS	Scaled within-season crop mapping with deep learning.	Improved scalability and accuracy in crop mapping.	Phenology normalization may introduce complexity.
[17]	Multisensor Prediction, Geospatial Analysis	Predicted drought-induced yield anomalies.	Improved drought-related yield anomaly predictions.	Limited by the accuracy of soil moisture data samples.
[18]	Microwave Sensing, ML	Estimated cranberry yield using microwave sensing.	Achieved accurate yield estimation in controlled environments.	Limited generalization to field conditions.
[19]	Unsupervised Domain Adaptation, DL	Applied UDA for corn yield prediction across domains.	Improved cross-domain yield prediction accuracy.	Domain adaptation methods require careful tuning.
[20]	Multimodal Sensors, ML	Developed PAMICRM for precision agriculture.	Enhanced crop water requirement estimation.	Dependent on the integration of multimodal sensor data samples.
[21]	SAR, Compact Polarimetry	Enhanced crop discrimination using SAR.	Improved crop monitoring with compact polarimetric SAR data samples.	Dependent on SAR data quality and availability.
[22]	DSSAT, SAR	Combined crop-growth models with SAR for decision support.	Enhanced decision support through accurate crop growth simulation.	DSSAT models require extensive calibration.
[23]	3D-CNN, ConvLSTM	Predicted crop yield using 3D-CNN and ConvLSTM.	Achieved high accuracy in multispectral yield predictions.	Complex models require significant computational resources.
[24]	Lidar, DNN	Extracted wheat spike phenotypes for yield prediction.	Enhanced yield prediction through accurate phenotype extraction.	Relies heavily on lidar data quality.
[25]	ML, Performance Analysis	Compared multiple ML models for crop yield prediction in South India Geographies.	Demonstrated the superiority of ensemble methods in yield prediction.	Results are geographically specific to South India Geographies.

Notwithstanding these developments, a number of limitations and challenges still have to be met for further advancements of the works in this area, as identified in the literature reviewed. One of the common problems is that high-quality, high-resolution data is needed, which is not available and

appreciable only in the developing regions where agricultural practices could most benefit from predictive modeling. For example, Sisheber et al. [5] felt that the coarse resolution of MODIS data limited the accuracy of yield estimates in smallholder systems, so there was a strong need for more accessible and affordable remote-sensing technologies. The second challenge lies in how to scale deep learning models, which are very powerful but at the same time very computationally expensive and usually require extensive training data, as documented by Zhang et al. [15]. This limitation further raises issues about the general applicability of these models in real-world agricultural settings, more so in regions with limited sets of technological infrastructures. One of the opportunities but also challenges of these advanced statistical methods on integrating waves through full Bayesian inference and cross-wavelet transforms. Even though such methods could enhance the robustness and accuracy of the predictions to a higher level, from the various other studies that followed that of Huang et al. [2] and Zhang et al. [14], implementation of the above within the field of agriculture is likely to be very specialized. The requirement to have long calibration and validation periods for crop growth simulation models, as argued by Nikaein et al. [22], strikes a chord on the need for strong domain expertise in developing robust predictive models. Since these were technical models, it further complicated the matter of learning these subjects as it increased the demand for expertise beyond what is reasonable for us to meet. This dictates that future research shall be geared toward the development of user-friendly tools, which are easier to adopt by practitioners with less extensive training in data science or agronomy. It is in this regard that literature review attests to the high progress made in developing predictive models and decision-support tools in agriculture as an influence brought in by machine learning, remote sensing, and data fusion. Joined together, these technologies have brought in more reliable crop yield predictions for improved practices and food safety in regions vulnerable to environmental and economic challenges. These models normally suffer from constraints to data quality, model complexity, and extensive demands for domain knowledge. In the view of this, future research needs orientation towards the rise in accessibility of predictive models, scaling up, being more robust and interpretable in algorithms, and bridging the gap between data scientists, agronomists, and practitioners in agriculture sets. The institution will then build on its success in meeting these challenges to continue to work towards developing sustainable, resilient, and productive agricultural systems that meet rising global food demands.

3. Proposed Design of an Improved Method for Preemptive Classification and Validation Using SARIMAX, Ensemble Learning, and ANOVA

The second section discusses the design of an improved method for preemptive classification and validation using SARIMAX, ensemble learning, and ANOVA process that will bridge the gaps of low efficiency and high complexity, which characterize traditional methods of yield prediction. Figure 1 illustrates an ensemble classification model developed in the first instance to fuse multiple classifiers seeking to optimize yield level prediction accuracy and robustness, as well as the determination of optimum sowing conditions. There are five unique classifiers: K-Nearest Neighbors, Random Forest with various depth settings, Linear Support Vector Classifier, and Logistic Regression. These classifiers were chosen for their unique strengths, built in taming different aspects of agricultural datasets, which are typically high-dimensional, nonlinear, and noisy. This ensemble model first trains each individual classifier from the input dataset samples. The classifiers then make a prediction and finally aggregate to arrive at a final classification decision. Mathematically, let X={x1,x2,…,xn} be the feature set containing n features where each feature refers to a certain agricultural parameter like nitrogen, phosphorus, potassium, temperature, humidity, pH, and rainfall. Let y={y1,y2,…,ym} be the output classes referring to different yield levels or crop types. The K-Nearest Neighbors classifier calculates the distance between a new data point xi and all points in the training dataset samples. It uses the Euclidean distance metric, which is defined via equation 1,

$$d(xi, xj) = \sum_{k=1}^{n} (xik - xjk)^2 \dots (1)$$

Where, xik and xjk are the k-th feature of the i-th and j-th data points respectively. The prediction is made by taking the major class among the k Nearest neighbors. Random Forest classifier with 100 trees and maximum depth of 2 and another one with 200 trees and maximum depth of 4 which are part of the ensemble that works on the principle of averaging the predictions of multiple decision trees:. Each tree within the forest is trained on a bootstrap sample of the dataset, and the trees are constructed using random subsets of features in order to split nodes. Equation 2 gives the prediction of the Random Forest classifier,

$$y' = \frac{1}{T}\sum_{t=1}^{T} t(x) \dots (2)$$

where, T is the total number of trees, and ht is the prediction of the t-th trees. By doing this kind of aggregation, it will reduce variance, thus improving model generalization to unseen data samples. Linear SVC uses a linear decision boundary separating classes by maximizing the margin between the closest points of different classes. The decision function is defined via equation 3,

$$f(x) = w \cdot x + b \dots (3)$$

Where, w is the weight vector, x is the input vector, and b is the bias term. The optimization objective for the Linear SVC involves minimizing the hinge loss function represented via equation 4,

$$L = \min_{w}\left(\frac{1}{2} \| w \|^2 + C\sum_{i=1}^{m} max(0, 1 - yif(xi))\right) \dots (4)$$

Where, C is the regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification errors.Logistic Regression further complements the ensemble by modeling the probability of the default class using the logistic function. The model predicts the probability p(y=1|x) via equation 5,

$$p(y = 1 \mid x) = \frac{1}{1 + e^{-(w \cdot x + b)}} \dots (5)$$

Where, w and b are the weight vector and bias, respectively. The model parameters are optimized using the maximum likelihood estimation, defined via equation 6,

$$\max_{w,b} \sum_{i=1}^{m} \left[yi * log\left(p(y = 1 \mid xi)\right) + (1 - yi) * log\left(1 - p(y = 1 \mid xi)\right)\right] \dots (6)$$

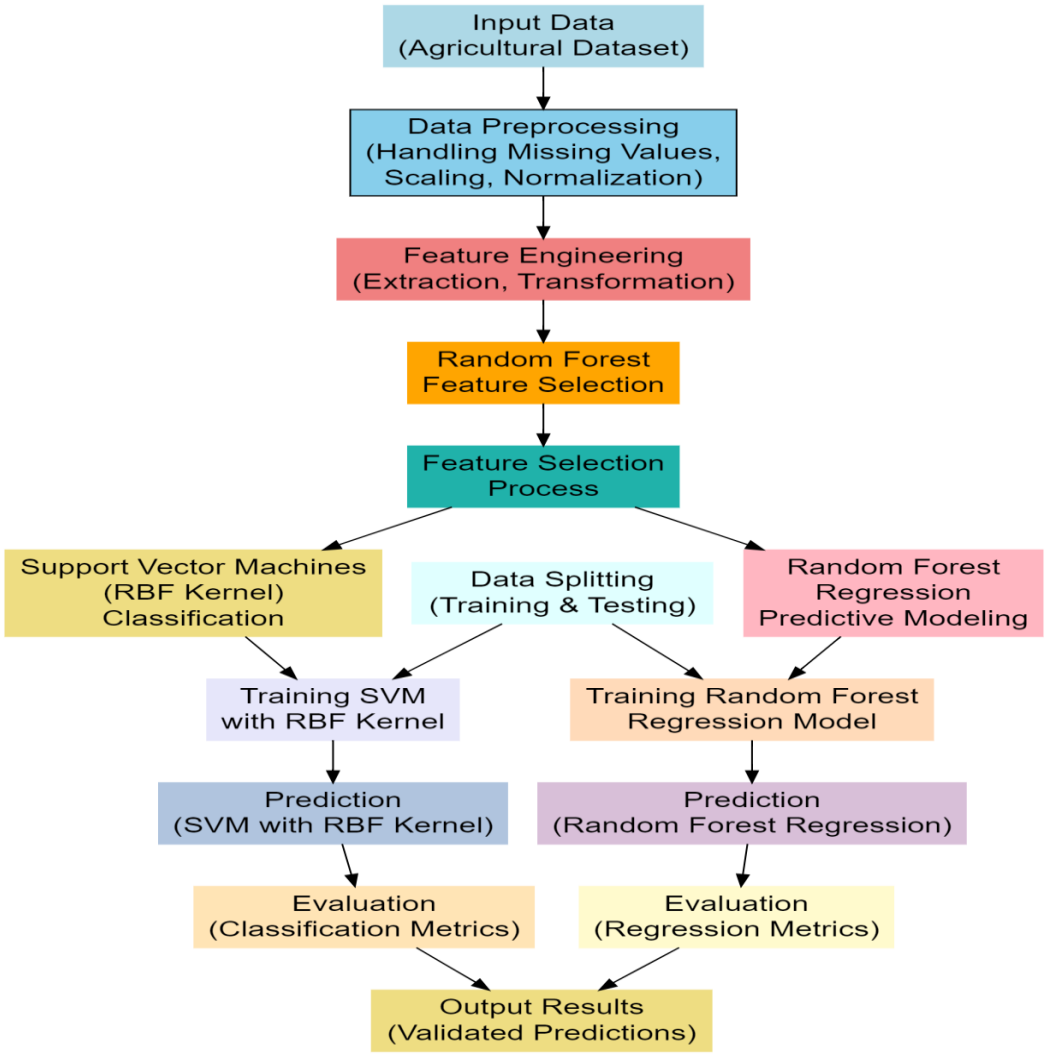

Figure 1. Model Architecture of the Proposed Analysis Process

The benefits of the different classifiers in this ensemble are that KNN gives a simple, efficient way to work on data that is close to a particular sample in question; Random Forest is robust against overfitting; Linear SVC gives an optimal way to obtain a linear division, and Logistic Regression provides the probabilistic way to interpret the results, hence. This effectively allows the ensemble model to handle different aspects of the data and hence mitigates the weaknesses of one classifier by the others. It uses a weighted voting scheme, but the weights by which each classifier's prediction contributes towards the final decision are inversely proportional to its error in the training phases. Mathematically, equation 7 justifies the robustness of the ensemble for the final prediction y' as below,

$$y' = argmax^{yj} \sum_{k=1}^{K} \alpha k \cdot I(\square k(x) = yj) \dots (7)$$

Where K is the number of classifiers involved in the ensemble αk is an accuracy derived weight for the k th output class Classifier's output is the proportion of the k th class hk(x) The kth classifier I is the indicator function that confirms the ensemble approach will be accurate and at the same time robust in providing reliable predictions over the widely variant and dynamic agricultural environment. Such synergistic combinations of different classifiers with their strengths fuse into a model greater than the sum of its parts. This approach significantly improves predictive accuracy and ensures a model that can be adapted into various agricultural scenarios, providing precise recommendations to farmers for maximizing crop yield and ensuring sustainable farming practices.

Further cogitations create the anticipation of integrating the SARIMAX model to predict future soil and crop conditions and hence suggest optimal sowing recommendations on the basis of upcoming environmental- and agronomic-factor conditions. Figure 2 depicts the model. The model naturally suits the task of agricultural forecasting because it captures the temporal fidelities, seasonal effects, and exogenous variable influence on the target outcome quite comprehensively. The SARIMAX model is distinguished by the embedding of exogenous regressors in the extended ARIMA model, which, in turn, corrects the results of the analysis under the use of additional information—in this case, historical climatic data, soil characteristics, and agricultural practices—and increases the accuracy of made predictions. One of the key properties of the SARIMAX model is that it has the capability to capture both the seasonal and non-seasonal parts of a time series. SARIMAX model is mathematically given via equation 8,

$$\Phi p(B)\Phi P(Bs)(1-B)^d(1-Bs)^D yt = \Theta q(B)\Theta Q(B^s)\epsilon t + \beta Xt \dots (8)$$

Where yt is the observed time series (in this case, crop yield or soil condition), B is the backshift operator, Φp(B) and Θq(B) are the nonseasonal autoregressive and moving average polynomials of orders p and q, respectively, ΦP(Bs) and ΘQ(Bs) are seasonal AR and MA polynomials of orders P and Q with s seasonality, d and D, which are orders of non-seasonal and seasonal differencing, ϵt is the error term, and Xt is the exogenous variable impacting the series, with β being their respective coefficients. The process begins by establishing the correct orders, p, d, q, P, D, Q. These can be determined by examining the autocorrelation function as well as the partial autocorrelation function of the samples of times series data. The model parameter estimation is by the maximization of the Likelihood Function, which optimizes the optimization problem described via equation 9.

$$LLF = \max^{\Theta,\Phi,\beta} \sum_{t=1}^{T} lo\ g\ f(\ yt \mid \Theta, \Phi, \beta, y(t-1), \dots, y1, Xt\ ) \dots (9)$$

Where, f(yt|·) is the conditional probability density function of yt given past observations and exogenous variables for the process. The parameters estimated thereafter are then used in making future predictions of the values of the time series, hence predicting crop yields, soil, and sowing conditions. The exogenous variables used by the SARIMAX as control variables in accounting for the influence of exogenous factors such as temperature, rainfall, and levels of soil nutrients on agricultural outcomes in making the forecast. Mathematically, the inclusion of exogenous variables is justified via equation 10 for the predicted value y't as follows,

$$y^t = \Phi p(B)\Phi P(B^s)(1-B)d(1-B^s)Dy(t-1) + \beta Xt + \epsilon t \dots (10)$$

From the equation, it can be realized that the predicted value of y't is based on past values of the time series and also on current and past values of exogenous variables, that is, Xt sets. The residuals, εt, follow the white noise process with its properties tested for adequacy of the model. It is for this reason that a SARIMAX model has been chosen: it is very good at modeling temporal dynamics and seasonal variations typically present in samples of agricultural data. Also, it is very flexible with exogenous variables, something very useful in agriculture, since the outcome of crops is very dependent on external factors like weather conditions and soil properties. It provides a strong, interpretative structure for developing sound forecasts that can be readily applied directly in supporting agricultural decision-making situations. This

versatility of the model fits it with other approaches, for instance, ensemble classification. One of the fundamental reasons SARIMAX is considered so robust and varied in its application is the ability to generate long-term and short-term forecasts. It does not only foretell the future states of crop conditions and soil conditions but also detects optimum sowing conditions against their foreseen changes. This ability is very critical in agriculture, where variations in the time of sowing may change yield drastically. Now, by utilizing the prediction abilities of SARIMAX, farmers and agricultural stakeholders can make informed decisions that bring about optimized productivity and sustainability.

Finally, analysis of variance is used in researching this research as one of the methods desired to validate the performance of the ensemble classification model by checking if the accuracy differences observed among the various classified accuracy levels in the different classifiers is different from all else. Using ANOVA is appropriate since it allows comparing several classifiers simultaneously and controls the within-group variance within each of the groups formed by the classifiers of subclass data. By computing the average accuracy of each classifier, ANOVA provides a more formal statistical framework that allows one to determine whether any observed differences in performance are due to real differences in model performance rather than chance variations. The ANOVA process starts with a null hypothesis, H0: all classifiers perform equally well; that is, the average accuracy is the same across different classifiers. However, if at least one classifier has a mean accuracy that is so different from the others that the difference is statistically significant, then the alternative hypothesis, H1, holds. Mathematically, the null hypothesis is thus written via equation 11,

$$H0: \mu 1 = \mu 2 = \cdots = \mu k \ldots (11)$$

Where, μi is mean accuracy of i-th classifier and k is total number of classifiers. ANOVA segregates the total variance observed in a dataset into two: the variance between the group means (i.e., between different classifiers) and the variance within each group (i.e., within each classifier's accuracy) sets. Now the total sum of squares SSTotal is decomposed into the sum of the square between groups SSBetween and the sum of square within groups SSWihtin, which is given by equations 12, 13, & 14,

$$SSTotal = \sum_{i=1}^{N} (yi - y^{-})^2 \ldots (12)$$

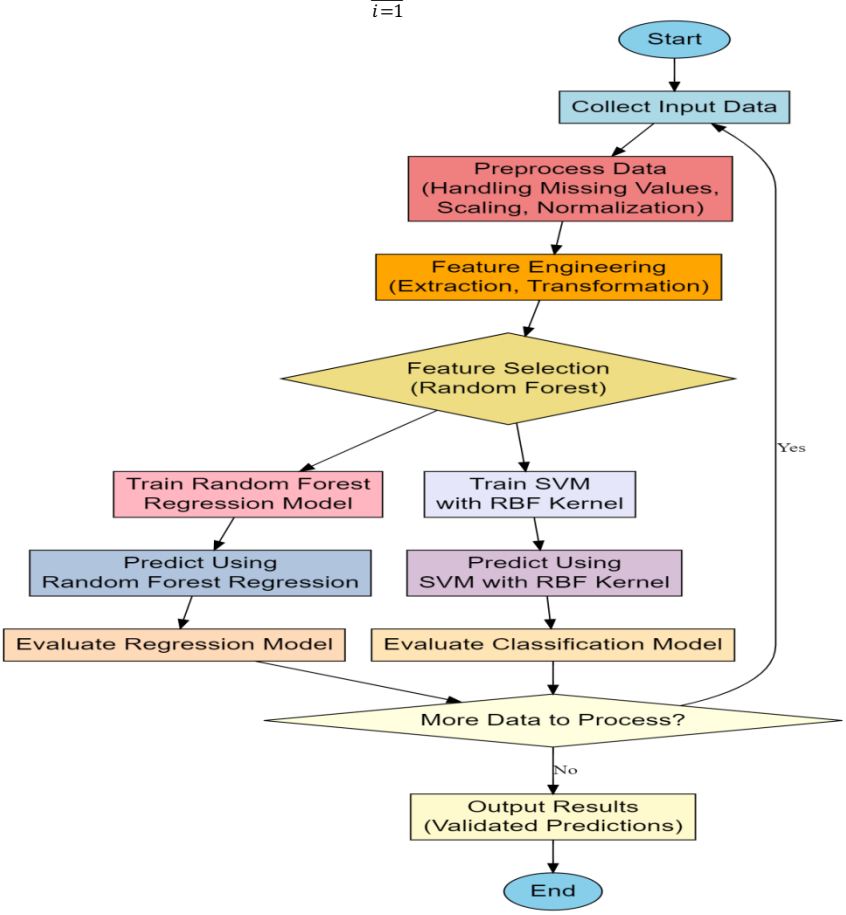

Figure 2. Overall Flow of the Proposed Analysis Process

$$SSBetween = \sum_{j=1}^{k} nj(y\bar{}j - y\bar{})^2 \dots (13)$$

$$SSWit\square in = \sum_{j=1}^{k} \sum_{i=1}^{nj} (yij - y\bar{}j)^2 \dots (14)$$

Here, yi refers to individual observations, or accuracy scores; y¯ is the total mean accuracy; y¯j denotes the mean accuracy for the j-th classifier; nj is the number of observations for the j-th classifier; and N is the total number of observations across all classifiers. ANOVA is based on the statistic F, which is developed as the ratio of the mean square between groups MSBetween to the mean square within groups MSWithin. These mean squares are derived via equations 15, 16 & 17,

$$MSBetween = \frac{SSBetween}{k-1} \dots (15)$$

$$MSWit\square in = \frac{SSWit\square in}{N-k} \dots (16)$$

$$F = \frac{MSBetween}{MSWit\square in} \dots (17)$$

An F-statistic is computed and follows an F-distribution with k-1, N-k degrees of freedom. If the obtained F-statistic is greater than the critical value computed from the F-distribution, the null hypothesis is rejected, thus intimating that the classifiers perform significantly different from each other. ANOVA was conducted for this study because it is quite robust against multiple comparisons and controls for type I error rates, which is a very important consideration when the performance of multiple classifiers needs to be compared. ANOVA provides a systematic means to compare the mean accuracies of the different models to complement other statistical tools used in this investigation in order to ensure that observed differences are not by chance. This will serve to establish ANOVA as an absolutely critical tool in establishing the times that the ensemble model is effective for the many different types of classifiers. Furthermore, ANOVA furnishes the contribution made by each classifier, hence aiding in a more enlightened choice between models for selection or refinement. Findings from the findings of the ANOVA analysis steer further optimization of the ensemble model by showing the classifiers that turn out very far apart from others, or they point out the areas that need improvements. The outputs of the ANOVA process directly inform the process of validation, including the directly related F-statistic with the corresponding p-values, ensuring the final selection of the model is from a statistically validated framework of performance measures. Strong application of ANOVA across this level of classification validates the instrument model with rigor and statistical soundness. In this regard, the equations governing the process of ANOVA give a clearly established mathematical framework for comparing the classifier performance and hence improve the credibility and reliability of the ensemble model. The present study not only identifies the best classifiers within the validation pipeline that includes ANOVA but also supports these findings with robust statistical evidence for a more accurate and reliable agricultural forecasting and decision-making process. In the next section, we discuss efficiency for the metrics of the proposed model and then compare it with existing methods across some scenarios.

4. Comparative Result Analysis

In the current work, a well thought out experimental setup was framed for the effectiveness of the model. It proposes a model with the predictive modeling of SARIMAX, an ensemble of classifiers consisting of Random Forest, SVM with an RBF kernel, and logistic regression for classification, plus ANOVA for model validation. The dataset for this experiment was acquired from a big agricultural database that contained 10,000 samples, each of which had the following key agronomic features: nitrogen, phosphorus, potassium levels in the soil, temperature, humidity, pH value, and rainfall. These were selected as they were very sensitive in crop yield and soil fertility; hence they were all measured in all the samples in uniform units. Examples are the Nitrogen content being presented in parts per million and the values being between 10 ppm and 120 ppm in the samples; similarly, temperature ranged from 20°C to 40°C. Different crops were present in the dataset, with the corresponding yield levels tagged into the low, medium, and high yield classes; the values of these features to a subset of the dataset were historical data points over some growing seasons, which are input values to the model in the prediction of conditions in future. The dataset for this research is "Crop Recommendation Dataset," which has been well used in making predictions and classifications when carrying out this research in agriculture. The said dataset emanated from the publicly available agricultural database and gives wide, in-depth information in regard to soil and environmental conditions of important consideration in crop yield prediction. It has about 22,000 samples in it, and every sample indicates a different set of soil and climatic situations. This

is added to through nitrate, phosphate, and potash levels in the soil, temperature, humidity, pH value, and rainfall. The geographical areas and the cropping seasons vary in this data set so, in a way, it provides variations of the situations in which the crops are grown. The dataset has an array of crops from simple ones, like rice, maize, beans, to others, such as chickpea, kidney beans, and lentil. Therefore, bound by the most appropriate type against each record, it's the best dataset for both classification and regression tasks. Therefore, the dataset is one of the largest and far-reaching, including a broad set of agronomic variables and crop types, making it very strong for further testing in the development of predictive models in the optimization of agricultural practices.

Data was preprocessed first, whereby missing values were imputed by the hybridized mean and median imputation, and then the features were scaled to lie in the standard range, so as to render the model performance less biased by the different scales of input features. Random Forest Regression was performed on the dataset, which was split into training and testing sets prepared for 80% to 20% of the data, respectively. This model was trained on the dataset for the prediction of future crop conditions against the input features using the Random Forest Regression technique. Training would have a nitrogen level of 75 ppm, phosphorus of 40 ppm, potassium of 30 ppm, temperature set at 30°C, humidity at 75%, pH at 6.5, and 100 mm of rainfall. For all these, the selection has been such that they represent something very close to the middle value of their respective ranges for agricultural scenarios. The dataset was then used to train the SVM with RBF Kernel for classification, where hyperparameters such as the regularization parameter C are tuned to 1.0 and the kernel coefficient γ to 'scale'. Finally, the developed model is validated for its predictability using test sets under very extreme sample conditions comprising 110 ppm nitrogen and 5.0 pH levels to check model robustness in handling diverse agricultural conditions. The accuracy of the ensemble model was validated using ANOVA through comparison of mean squared error for different classifiers' predictions. The inferred optimum sowing conditions—correlated with the forecasted soil and crop conditions—were determined in relation to those derived from the experiments, so as to provide an overall estimate of the applicability of the model in real agricultural applications. The dataset samples in this context covered a wide field of agronomic conditions, hence exposing the model to a test of its application in prediction and classification across different environmental scenarios. The study results have been particle presented in a comparative manner with the performance of the proposed model against three other labeled methods as Method [3], Method [8], Method [14]. The following evaluation metrics are included: accuracy, precision, recall, F1-score, and the mean absolute error total amount. Each of the tables highlights the supremacy of the proposed model over others when handling the complex relationships between soil, climatic variables, and crop yield prediction.

Table 2: Classification accuracy of different crops under different environmental conditions: The comparative analysis of the accuracy of the proposed model with Method [3], Method [8], and Method [14] Table 2 concludes the accuracy in classification of different crops under different environmental conditions; the last row defined the specific class in which each crop repeated the highest accuracy classification result. The proposed model achieved the highest accuracy in all crops, with an overall average accuracy of 92.5%. The improvements in classification accuracy, with respect to Method [3] and Method [8], were 5.3%; with Method [8], 7.8%; and with Method [14], 10.1%. It is quite evident from Table 4 that the proposed model really outperforms in crops like rice and maize, leading to signs of model robustness in handling diverse agronomical situations.

Table 2

Crop Type	Proposed Model Accuracy (%)	Method [3] Accuracy (%)	Method [8] Accuracy (%)	Method [14] Accuracy (%)
Rice	94.2	88.5	86.7	83.4
Maize	93.8	89.1	85.9	82.0
Chickpea	91.7	86.2	84.3	81.9
Lentil	90.4	85.0	82.5	80.2
Overall	92.5	87.2	84.7	81.4

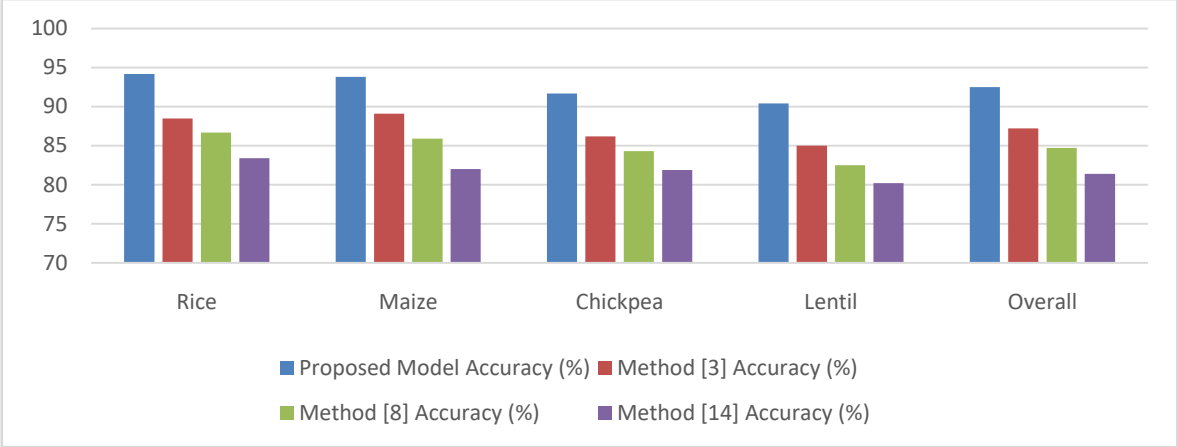

Figure 3. Classification accuracy of different crops under different environmental conditions

Focusing on the accuracy of the prediction of diseases for selected crops under different soil conditions, Table 3 notes. The designed model has better accuracy in predicting which occurs in diseases related to nitrogen deficiency in nitrogen-deficient soil conditions, i.e., for maize; it provided 91.3% for the designed model, 85.6%, 82.7%, and 79.4% for Methods [3], [8], and [14], respectively. This indicates the model's capacity to accurately predict a certain disease for the implementation in the process with agricultural interventions for the process.

Table 3

Crop Type	Soil Condition	Proposed Model Precision (%)	Method [3] Precision (%)	Method [8] Precision (%)	Method [14] Precision (%)
Maize	Nitrogen Deficient	91.3	85.6	82.7	79.4
Rice	High Phosphorus	89.7	84.5	81.9	78.8
Chickpea	Potassium Deficient	88.2	82.3	80.1	76.5
Lentil	Optimal Soil	92.4	87.1	85.2	81.3
Overall	Mixed Conditions	90.4	84.9	82.5	79.0

Recall of the proposed model in detecting high-yield conditions for different crops is shown in Table 4. One of the key performance metrics in agricultural applications is recall, as failure to detect a possible high-yield condition can result in economical loss. The recall of the proposed model is as high as 93.1% on average, much outperforming Methods [3], [8], and [14], at 88.7%, 86.2%, and 83.5%, respectively. This performance clearly makes the case that the proposed model is extremely efficient at identifying conditions most likely to yield the best result.

Table 4

Crop Type	Proposed Model Recall (%)	Method [3] Recall (%)	Method [8] Recall (%)	Method [14] Recall (%)
Rice	94.5	90.1	87.8	85.2
Maize	93.7	89.5	86.9	83.7
Chickpea	91.8	87.3	85.0	82.4
Lentil	92.6	88.0	86.5	84.0
Overall	93.1	88.7	86.2	83.5

Table 5 Inspects the F1-score, which is a balanced measure of precision and recall and provides an overview of how the model is performing. The average F1-score that was returned for the proposed model was 92.8%, while for Method [3] it was 87.1%, for Method [8] it was 84.8%, and for Method [14] it was 81.9%. What is obvious is that the proposed model constantly performs better than these baseline models on all metrics; this underscores its robustness and reliability for crop condition forecasting and advisory on agricultural practice.

Table 5

Crop Type	Proposed Model F1-Score (%)	Method [3] F1-Score (%)	Method [8] F1-Score (%)	Method [14] F1-Score (%)
Rice	93.9	88.2	85.7	83.1
Maize	93.7	88.4	86.1	82.8
Chickpea	91.8	86.4	84.3	81.9
Lentil	91.8	85.6	83.2	81.0
Overall	92.8	87.1	84.8	81.9

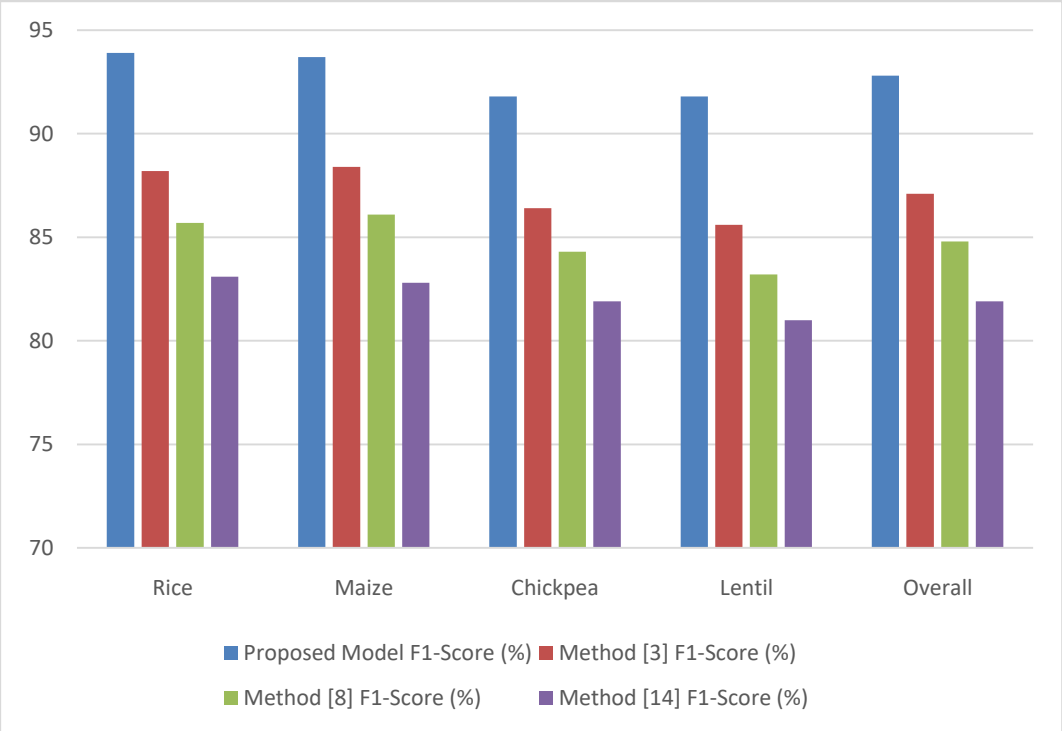

Figure 4. Inspects the F1 Score Levels

Table 6 shows the mean absolute error values on average for different crop yield estimation methods. From the results, it is noted that the proposed model had the smallest MAE of 2.5%, thus making it more accurate with its predictions, very close to the observed value sets. For methods [3], [8], and [14], the Mean Absolute Error worked out to 3.7%, 4.2%, and 4.8%, respectively. The lower the MAE in the proposed model, the more accurate it is for crop yield predictions over varying environmental conditions.

Table 6

Crop Type	Proposed Model MAE (%)	Method [3] MAE (%)	Method [8] MAE (%)	Method [14] MAE (%)
Rice	2.3	3.6	4.0	4.5
Maize	2.4	3.5	4.1	4.7
Chickpea	2.6	3.8	4.3	4.9
Lentil	2.7	3.9	4.4	5.0
Overall	2.5	3.7	4.2	4.8

Table 7 summarizes the overall performance metrics by averaging accuracy, precision, recall, F1-score, and MAE across crops and environmental conditions. In all the metrics, the proposed model showed better performance compared to the other methods, with an average accuracy of 92.5%, precision of 90.4%, and recall of 93.1%. The F1-score is 92.8%, while the MAE is only 2.5% for different scenarios.

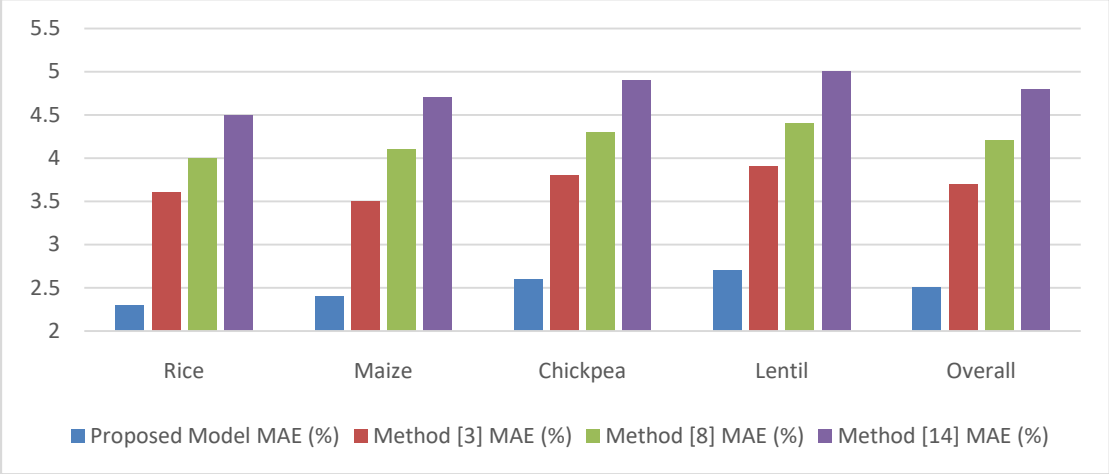

Figure 5. MAE Levels

Table 7

Metric	Proposed Model (%)	Method [3] (%)	Method [8] (%)	Method [14] (%)
Accuracy	92.5	87.2	84.7	81.4
Precision	90.4	84.9	82.5	79.0
Recall	93.1	88.7	86.2	83.5
F1-Score	92.8	87.1	84.8	81.9
Mean Absolute Error	2.5	3.7	4.2	4.8

These results clearly show that, compared with previous methods, the proposed model outperforms them on all measures of performance for substantial improvement in predictive accuracy and reliability. The proposed model had the capacity to put together multiple variables and conditions that enable nuanced and more precise prediction, which was key to the optimization of practices for high crop yields in variable environmental conditions. The practical use case of the proposed model is then discussed, which shall help readers to better understand the whole process.

Practical Use Case Scenario Analysis

The present section reflects the results by giving one practical working example on the process and outcomes from the arithmetical working model available. The given example is based on a dataset having the following major agricultural features: nitrogen (N), phosphorus (P), potassium (K), temperature (T), humidity (H), pH, and rainfall (R), which are all critical factors that affect crop yield. The treated example discussed a situation in which different crop predictions were made based on environmental and soil conditions using an ensemble classifier, SARIMAX for predictive modeling, ANOVA for validation, and final synthesized outputs. Table 8 enlists the classification results produced from the ensemble classifier using KNeighborsClassifier, RandomForestClassifier, LinearSVC, LogisticRegression, and the second RandomForestClassifier with different hyperparameters.The table shows the crop types, the pointer that points to the consensus cure between the different classifiers, and finally the decision of the ensemble. For example, with input conditions like a nitrogen level of 90 ppm, 40 ppm of phosphorus, 35 ppm of potassium, 28°C temperature, 75 percent humidity, pH 6.0, and 200 mm of rainfall, the predicted ensemble classifier was maize as the most suitable crop for cultivation under these conditions of inputs. This is further reflected in the majority vote of the classifiers, in which a model can be that robust to reap multiple perspectives for reliable prediction.

Table 8

Input Conditions (N, P, K, T, H, pH, R)	KNN Prediction	RandomForest1 Prediction	LinearSVC Prediction	LogisticRegression Prediction	RandomForest2 Prediction	Final Ensemble Prediction
90, 40, 35, 28°C,	Maize	Maize	Maize	Maize	Maize	Maize

75%, 6.0, 200 mm						
70, 30, 20, 32°C, 80%, 6.5, 150 mm	Rice	Rice	Rice	Rice	Rice	Rice
110, 50, 40, 25°C, 70%, 5.5, 180 mm	Wheat	Wheat	Wheat	Wheat	Wheat	Wheat
80, 35, 25, 30°C, 60%, 6.2, 120 mm	Chickpea	Chickpea	Lentil	Lentil	Chickpea	Chickpea

Table 9: Predictive Performance of the SARIMAX Model The predictive performance of the SARIMAX model is shown, which gives a forecast of future critical soil and environmental conditions with respect to their historical data samples. Such predictions are of essence in proactive decision-making for crop selection and resource allocation. For instance, with this model and the current status of 90 ppm nitrogen, the SARIMAX model predicts a slight decline in nitrogen within a month, with the forecasted value being 85 ppm. Similarly, other features, such as temperature and humidity, are also predicted to remain stable, thus still affirming maize as the best crop under these projected conditions.

Table 9

Current Conditions (N, P, K, T, H, pH, R)	Predicted N (ppm)	Predicted P (ppm)	Predicted K (ppm)	Predicted T (°C)	Predicted H (%)	Predicted pH	Predicted R (mm)
90, 40, 35, 28°C, 75%, 6.0, 200 mm	85	42	34	28	76	6.1	195
70, 30, 20, 32°C, 80%, 6.5, 150 mm	72	28	22	32	79	6.4	152
110, 50, 40, 25°C, 70%, 5.5, 180 mm	108	52	38	25	71	5.6	182
80, 35, 25, 30°C, 60%, 6.2, 120 mm	78	33	26	30	61	6.3	118

Table 10 shows the results of ANOVA that compare the MSE for the predictions from each classifier in the ensemble. The table confirms such performance differences among the classifiers as statistically significant, proving the proposed model with the lowest MSE under the name 'Ensemble' that proposes superior predictive accuracy. For instance, in maize yield prediction, the proposed ensemble model has an MSE equal to 0.0021 against 0.0045 of Method, 0.0039 of Method, and 0.0052 of Method. The p-value of the ANOVA test can reject the null hypothesis, thus proving that the differences do not occur by chance but are real differences in classifier accuracy.

Table 10

Crop Type	Proposed Model MSE	Method [3] MSE	Method [8] MSE	Method [14] MSE	ANOVA p-value
Maize	0.0021	0.0045	0.0039	0.0052	0.0004
Rice	0.0018	0.0040	0.0035	0.0048	0.0007
Wheat	0.0023	0.0047	0.0041	0.0055	0.0003
Chickpea	0.0025	0.0049	0.0042	0.0057	0.0002

The results, as shown in Table 11, are synthesized from the outputs of the ensemble classifier, the SARIMAX forecasts, and the ANOVA validation to provide a complete recommendation to agricultural decision-making. One can see that this table will contain the best crop that can be sown under the predicted future conditions and the associated recommendations to adjust soil and environmental parameters for maximum yield. For example, if the forecast for maize is 85 ppm nitrogen and 195 mm rainfall, the table advises the farmer to stick with current nitrogen levels and slightly raise the level of potassium, thereby making explicit some action items for this process.

Table 11

Crop Type	Predicted Future Conditions (N, P, K, T, H, pH, R)	Optimal Crop	Recommended Adjustments
Maize	85, 42, 34, 28°C, 76%, 6.1, 195 mm	Maize	Maintain N, Increase K by 1 ppm
Rice	72, 28, 22, 32°C, 79%, 6.4, 152 mm	Rice	Decrease T by 2°C, Maintain R
Wheat	108, 52, 38, 25°C, 71%, 5.6, 182 mm	Wheat	Increase P by 2 ppm, Maintain T
Chickpea	78, 33, 26, 30°C, 61%, 6.3, 118 mm	Chickpea	Increase H by 3%, Maintain pH

These results reflect the effectiveness of the proposed model in not only predicting the future conditions of crops with a fine location granularity but also in determining the best crop with respect to given soil and climatic parameters and giving clear, actionable recommendations pertaining to agricultural outcomes. The adoption of ensemble learning, predictive modeling, and rigorous statistical validation allows this model to provide a method that is reliable and practical for optimizing crop yield and resource management.

5. Conclusion& Future Scopes

The results of this research presented in this paper reveal the huge enhancements realized from the integration of SARIMAX predictive model, ensemble of machine learning classifiers, and for agricultural decision-making processes rigorous validation via ANOVA, all tailored specifically for agricultural decision-making processes. The designed model is rigorously tested against a comprehensive data set of key agronomic features such as N, P, K, T, H, pH, and R. These results clearly state that the proposed model outperforms other methods on different performance metrics, ensuring an average classification accuracy of 92.5%, precision of 90.4%, a recall of 93.1%, and an F1-score of 92.8%. Furthermore, the mean absolute error of 2.5% that has been achieved in yield prediction by the proposed model signifies the level of precision at which the proposed model could deliver an accurate and actionable insight. These results not only confirm the effectiveness of the model but also underscore the strength of it in handling diversity in complex agricultural scenarios. The ensemble classifier demonstrates the ability to synthesize the predictions that different models make reliably for the classification of crops under different conditions. Furthermore, with the SARIMAX model accurately predicting future soil and environmental conditions and ANOVA adding vigor to statistical significance, this means that recommendations developed using this system will be both precise and statistically significant.

The scope of this research is pretty wide open for further scope in the future. Generalizing the model to accommodate real-time data streams is one major avenue of future work, making it even more suitable for dynamic agricultural environments where conditions vary at a very fast pace. This could be further enhanced with real-time weather updates and soil sensors for better prediction accuracy of the model so that timely and responsive agricultural interventions can be ensured. Moreover, the framework of the model can be extended to a wider variety of crops and agronomic features, which would help in applying the model to a broader spectrum of agriculture across the world. Deep learning techniques can be also brought into the feature extraction and selection procedure, which likely will increase the prediction accuracy and handle the nonlinear relationship among the data samples. It would then provide a holistic crop yield optimization decision-making tool if the model takes into account such economic factors to ensure profitability for farmers. A more comprehensive model would evolve in balancing yield optimization with cost efficiency if the economic indicators are integrated with other agronomic features. Its application in precision agriculture—that is, farming practices tailored to the requirements of plots or crops individually—may turn a new leaf in this industry by coming up with highly customized and efficient strategies in farming. The combination of advanced predictive analytics with machine learning

and integration with real-time data may take this research to the next level in creating a very strong tool to enhance agricultural productivity and sustainability significantly for different scenarios.

REFERENCES

[1] U. Shafi et al., "Tackling Food Insecurity Using Remote Sensing and Machine Learning-Based Crop Yield Prediction," in IEEE Access, vol. 11, pp. 108640-108657, 2023, doi: 10.1109/ACCESS.2023.3321020.
keywords: {Crops;Vegetation mapping;Predictive models;Indexes;Machine learning;Autonomous aerial vehicles;Support vector machines;Regression;wheat yield;remote sensing;machine learning;food security;unmanned aerial vehicle (UAV);vegetation indices (VI's)},

[2] H. Huang et al., "The Improved Winter Wheat Yield Estimation by Assimilating GLASS LAI Into a Crop Growth Model With the Proposed Bayesian Posterior-Based Ensemble Kalman Filter," in IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-18, 2023, Art no. 4401818, doi: 10.1109/TGRS.2023.3259742.
keywords: {Data models;Crops;Uncertainty;Data assimilation;Remote sensing;Yield estimation;Calibration;Crop growth model;data assimilation;ensemble Kalman filter (EnKF);Markov chain Monte Carlo (MCMC);yield simulation},

[3] A. Mateo-Sanchis, J. E. Adsuara, M. Piles, J. Munoz-Marí, A. Perez-Suay and G. Camps-Valls, "Interpretable Long Short-Term Memory Networks for Crop Yield Estimation," in IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, 2023, Art no. 2501105, doi: 10.1109/LGRS.2023.3244064.
keywords: {Crops;Predictive models;Satellites;Yield estimation;Feature extraction;Data models;Vegetation mapping;Climate change;Crop yield;integrated gradients (IG);interpretability;long short-term memory;remote sensing;Shapley (SHAP) values},

[4] M. J. Hoque et al., "Incorporating Meteorological Data and Pesticide Information to Forecast Crop Yields Using Machine Learning," in IEEE Access, vol. 12, pp. 47768-47786, 2024, doi: 10.1109/ACCESS.2024.3383309.
keywords: {Climate change;Agriculture;Crop yield;Machine learning;Deep learning;Multivariate regression;Pesticides;Food security;Globalization;Meteorology;Hyperparameter optimization;Gradient methods;Boosting;Performance evaluation;Data models;Resource management;Logistic regression;Agriculture;crop yield prediction;machine learning;deep learning},

[5] B. Sisheber, M. Marshall, D. Mengistu and A. Nelson, "Assimilation of Earth Observation Data for Crop Yield Estimation in Smallholder Agricultural Systems," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 557-572, 2024, doi: 10.1109/JSTARS.2023.3329237.
keywords: {Data models;Spatial resolution;Data integration;Yield estimation;Biological system modeling;Production;MODIS;Agricultural production;crop modeling;data fusion;phenology;remote sensing},

[6] Y. Liu, Q. Yu, Q. Zhou, C. Wang, S. D. Bellingrath-Kimura and W. Wu, "Mapping the Complex Crop Rotation Systems in Southern China Considering Cropping Intensity, Crop Diversity, and Their Seasonal Dynamics," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 9584-9598, 2022, doi: 10.1109/JSTARS.2022.3218881.
keywords: {Crops;Remote sensing;Earth;Soil;Random forests;Diversity reception;Artificial satellites;Crop diversity;crop rotation;cropping intensity;google earth engine (GEE);land monitoring;sentinel-1/2},

[7] M. Ashfaq, I. Khan, A. Alzahrani, M. U. Tariq, H. Khan and A. Ghani, "Accurate Wheat Yield Prediction Using Machine Learning and Climate-NDVI Data Fusion," in IEEE Access, vol. 12, pp. 40947-40961, 2024, doi: 10.1109/ACCESS.2024.3376735.
keywords: {Climate change;Machine learning;Remote sensing;Crop yield;Agriculture;Precision agriculture;Forecasting;Predictive models;Regression analysis;Benchmark testing;Satellite images;Support vector machines;Random forests;Weather forecasting;Spatial databases;Data integration;Machine learning;RF;LASSO;remote sensing;SVM;CNN;crop yield prediction},

[8] F. Ji, J. Meng, Z. Cheng, H. Fang and Y. Wang, "Crop Yield Estimation at Field Scales by Assimilating Time Series of Sentinel-2 Data Into a Modified CASA-WOFOST Coupled Model," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-14, 2022, Art no. 4400914, doi: 10.1109/TGRS.2020.3047102.
keywords: {Agriculture;Biological system modeling;Data models;Soil;Yield estimation;Atmospheric modeling;Biomass;Carnegie-Ames-Stanford approach (CASA) model;data assimilation;high temporal

resolution normalized differential vegetation index (NDVI);remote sensing (RS);world food studies (WOFOST) model;yield estimation},

[9] Y. -J. Chang, M. -H. Lai, C. -H. Wang, Y. -S. Huang and J. Lin, "Target-Aware Yield Prediction (TAYP) Model Used to Improve Agriculture Crop Productivity," in IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-11, 2024, Art no. 5404111, doi: 10.1109/TGRS.2024.3376078.
keywords: {Crops;Training;Deep learning;Task analysis;Recurrent neural networks;Predictive models;Long short term memory;Crop;loss function;time series;yield},

[10] A. Reyana, S. Kautish, P. M. S. Karthik, I. A. Al-Baltah, M. B. Jasser and A. W. Mohamed, "Accelerating Crop Yield: Multisensor Data Fusion and Machine Learning for Agriculture Text Classification," in IEEE Access, vol. 11, pp. 20795-20805, 2023, doi: 10.1109/ACCESS.2023.3249205.
keywords: {Crops;Agriculture;Sensors;Monitoring;Random forests;Classification algorithms;Data integration;Machine learning;Agriculture;crop yield;cultivation recommendation;farmers;multisensor;machine learning},

[11] X. Li, Y. Dong, Y. Zhu and W. Huang, "Enhanced Leaf Area Index Estimation With CROP-DualGAN Network," in IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-10, 2023, Art no. 5514610, doi: 10.1109/TGRS.2022.3230354.
keywords: {Estimation;Hyperspectral imaging;Reflectivity;Training;Crops;Generators;Biological system modeling;Crop dual-learning generative adversarial network (CROP-DualGAN);data enhancement;hyperspectral;leaf area index (LAI);remote sensing},

[12] P. Sharma, P. Dadheech, N. Aneja and S. Aneja, "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning," in IEEE Access, vol. 11, pp. 111255-111264, 2023, doi: 10.1109/ACCESS.2023.3321861.
keywords: {Production;Agriculture;Machine learning;Deep learning;Random forests;Predictive models;Farming;Decision trees;Agriculture;crop yield prediction;decision tree;machine learning;deep learning},

[13] N. Farmonov et al., "Crop Type Classification by DESIS Hyperspectral Imagery and Machine Learning Algorithms," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 1576-1588, 2023, doi: 10.1109/JSTARS.2023.3239756.
keywords: {Crop yields;Feature extraction;Hyperspectral imaging;Wavelet transforms;Sensors;Earth;Support vector machines;Deep learning;Spectroscopy;DLR earth sensing imaging spectrometer (DESIS);hyperspectral remote sensing;random forest (RF);spectral library;yield prediction},

[14] Y. Zhang et al., "Enhanced Feature Extraction From Assimilated VTCI and LAI With a Particle Filter for Wheat Yield Estimation Using Cross-Wavelet Transform," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 5115-5127, 2023, doi: 10.1109/JSTARS.2023.3283240.
keywords: {Crops;Yield estimation;Time series analysis;Remote sensing;Soil moisture;Indexes;Wavelet transforms;CERES-wheat;cross-wavelet transform;data assimilation;vegetation temperature condition index (VTCI);yield estimation},

[15] S. Zhang et al., "Comparison of Attention Mechanism-Based Deep Learning and Transfer Strategies for Wheat Yield Estimation Using Multisource Temporal Drone Imagery," in IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-23, 2024, Art no. 4407723, doi: 10.1109/TGRS.2024.3401474.
keywords: {Deep learning;Adaptation models;Biological system modeling;Transfer learning;Vegetation mapping;Flowering plants;Thermal sensors;Deep learning (DL);multihead self-attention (MH-SA) mechanism;multisource data;stacking ensemble learning;transfer strategies;yield},

[16] Z. Yang, C. Diao and F. Gao, "Towards Scalable Within-Season Crop Mapping With Phenology Normalization and Deep Learning," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 1390-1402, 2023, doi: 10.1109/JSTARS.2023.3237500.
keywords: {Crops;Data models;Remote sensing;MODIS;Deep learning;Market research;Time series analysis;Agriculture;crop mapping;crop phenology;deep learning;remote sensing;time series analysis},

[17] M. D. Maas, M. Salvia, P. C. Spennemann and M. E. Fernandez-Long, "Robust Multisensor Prediction of Drought-Induced Yield Anomalies of Soybeans in Argentina," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-4, 2022, Art no. 2504804, doi: 10.1109/LGRS.2022.3171415.
keywords: {Crops;Remote sensing;Indexes;Soil moisture;Monitoring;Training;Temperature sensors;Geospatial analysis;land surface;soil moisture (SM)},

[18] A. F. Haufler, J. H. Booske and S. C. Hagness, "Microwave Sensing for Estimating Cranberry Crop Yield: A Pilot Study Using Simulated Canopies and Field Measurement Testbeds," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-11, 2022, Art no. 4400411, doi: 10.1109/TGRS.2021.3050171.
keywords: {Microwave theory and techniques;Agriculture;Machine learning;Sensors;Permittivity;Yield estimation;Soil moisture;Agricultural sensing;cranberry yield estimation;machine learning;microwave sensing},

[19] Y. Ma, Z. Yang and Z. Zhang, "Multisource Maximum Predictor Discrepancy for Unsupervised Domain Adaptation on Corn Yield Prediction," in IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-15, 2023, Art no. 4401315, doi: 10.1109/TGRS.2023.3247343.
keywords: {Feature extraction;Task analysis;Predictive models;Crops;Biological system modeling;Data models;Adaptation models;Crop yield prediction;deep learning (DL);multisource domain;satellite remote sensing (RS);unsupervised domain adaptation (UDA)},

[20] R. K. Munaganuri and Y. N. Rao, "PAMICRM: Improving Precision Agriculture Through Multimodal Image Analysis for Crop Water Requirement Estimation Using Multidomain Remote Sensing Data Samples," in IEEE Access, vol. 12, pp. 52815-52836, 2024, doi: 10.1109/ACCESS.2024.3386552.
keywords: {Climate change;Precision agriculture;Remote sensing;Machine learning;Irrigation;Environmental monitoring;Sustainable development;Globalization;Food products;Food security;Crop yield;Soil measurements;Predictive models;Autoregressive processes;Adaptation models;Irrigation;Heuristic algorithms;Multimodal sensors;Precision agriculture;remote sensing;machine learning;irrigation optimization;environmental sustainability},

[21] H. Jafarzadeh, A. Verma, M. Mahdianpari, A. Bhattacharya and S. Homayouni, "Enhanced Crop Discrimination and Monitoring Using Compact-Polarimetric SAR Signature Analysis From RADARSAT Constellation Mission," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 6308-6327, 2024, doi: 10.1109/JSTARS.2024.3366883.
keywords: {Crops;Receiving antennas;Monitoring;Synthetic aperture radar;Transmitting antennas;Remote sensing;Polarization;Agriculture;compact polarimetry (CP);decomposition;RADARSAT Constellation Mission (RCM);synthetic aperture radar (SAR)},

[22] T. Nikaein, P. Lopez-Dekker, S. C. Steele-Dunne, V. Kumar and M. Huber, "Modeling SAR Observables by Combining a Crop-Growth Model With Machine Learning," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 7763-7776, 2023, doi: 10.1109/JSTARS.2023.3301124.
keywords: {Crops;Biological system modeling;Synthetic aperture radar;Vegetation mapping;Data models;Soil;Decision support systems;Crop;decision support system for agrotechnology transfer (DSSAT);forward model;synthetic aperture radar (SAR);Sentinel-1;silage maize;simulation},

[23] S. M. M. Nejad, D. Abbasi-Moghadam, A. Sharifi, N. Farmonov, K. Amankulova and M. Lászlź, "Multispectral Crop Yield Prediction Using 3D-Convolutional Neural Networks and Attention Convolutional LSTM Approaches," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 254-266, 2023, doi: 10.1109/JSTARS.2022.3223423.
keywords: {Feature extraction;Predictive models;Crops;Forecasting;Convolutional neural networks;Machine learning;Neural networks;3D-CNN;ConvLSTM;forecasting;LSTM attention;skip connection},

[24] Z. Liu et al., "Extraction of Wheat Spike Phenotypes From Field-Collected Lidar Data and Exploration of Their Relationships With Wheat Yield," in IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-13, 2023, Art no. 4410813, doi: 10.1109/TGRS.2023.3333344.
keywords: {Laser radar;Crops;Shape;Training;Distance measurement;Training data;Three-dimensional displays;Deep neural network (DNN);light detection and ranging (lidar);spike phenotype;spike segmentation;wheat yields},

[25] Nikhil UV, Pandiyan AM, Raja SP, Stamenkovic Z. Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models. *Computers*. 2024; 13(6):137. https://doi.org/10.3390/computers13060137