Linear Z Score and Gaussian Radial Artificial Neural Network Big Data Analytics to Enhance Crop Yield

C. V. Pallavi
Department of CSE, BNM Institute of Technology, Bangalore, India
pallavi.vasist@gmail.com (corresponding author)

S. Usha
Department of CSE, Rajarajeswari College of Engineering, Bangalore, India
usharesearch6@gmail.com

Received: 20 July 2024 | Revised: 31 July 2024 and 6 August 2024 | Accepted: 11 August 2024

 DOI: https://doi.org/10.48084/etasr.8442

ABSTRACT

Crop yield estimation is a pivotal matter in agricultural management, specifically under the backdrop of demographic growth and changing climatic conditions. Many studies have been conducted employing remote sensing for crop yield estimation. However, most were specifically concentrated on condition-based environmental monitoring systems. A shortage of exclusive applications persists regarding the use of remote sensing for soil health monitoring and implementing necessary measures to enhance crop yield. To address such insufficiency, the Linear Z-score and Gaussian Radial Artificial Neural Network-based (LZ-GRANN) crop yield estimation method is proposed in this paper to enhance productivity. The performance evaluation of the proposed LZ-GRANN method reduced the overall crop yield estimation time and error by 59% and 58% and improved precision and accuracy by 23% and 26% in comparison with the existing methods.

Keywords-crop yield estimation; linear mapping; standardized z-score; Gaussian Chebyshev; radial artificial neural network

I. INTRODUCTION

The two most paramount factors involved in crop monitoring and yield estimation are time and accuracy, directly influencing decision-making of agricultural policies and investments, as well as managing food enhancing the overall efficiency and market stability. Larger-scale crops monitoring is accelerated by remote sensing utilizing satellites, and Unmanned Aerial Vehicles (UAVs). Crop yield estimation models align with managing the challenges of agricultural production in accordance with productivity, environmental consequences and viability.

The multi-sensor Machine Learning Approach (MMLA) [1] is utilized to classify multisensory data valuable in the yield estimation. In the work herein, different machine learning algorithms such as decision tree, hoeffding tree, and random forest were employed and applied to multi-sensor data to improve precision and recall and reduce error rate significantly. However, the yield estimation error factor received less attention. In [2], Generative Adversarial Network (GAN) and Convolutional Neural Network (CNN) were utilized for data augmentation of remote sense and meteorological data [2]. First, the dimension of the input data was enlarged using GAN and accordingly the accuracy was measured. By using the combination of GAN and CNN the accuracy improved with minimum error while the crop yield estimation time was not focused. Crop yield estimation is a significant but complicated issue required for sufficient augmentation and well-organized utilization of natural resources. Crop yield estimations are exceptional to several stakeholders in the agri-food chains. A detailed performance analysis of monitoring agricultural resources assessing its accuracy and improvement for the European Union was investigated in [3]. Machine learning has been applied to propose a principle for large-scale crop yield estimation [4]. Nevertheless, the majority of research concentrates on the employment or implementation of domain application for classifying tasks like identifying crop type whereas the administration to regression tasks like crop yield prediction has been constrained. In [5-8], the accuracy was increased through Deep Neural Networks (DNNs) by employing three algorithms, namely a neural network using discriminative adversarial function, importance estimation pattern using Kullback-Leiblerand, and neural network employing transfer language [5-8].

II. RELATED WORKS

Considering the increasing demand for greater quantities of food, creating a precise mechanism to measure stress in crop

phenology and productivity is of greatest importance. Earth observation remote sensing data bestows a distinctive source of information to analyze and validate crops in a temporally resolved and spatially explicit fashion. In [9], the fusion of two novel techniques was employed via multisensor requipments for estimating crop yield [9]. A holistic review of crop yield estimation with the aid of machine learning techniques were investigated in detail in [10]. A systematic literature review using artificial intelligence in agriculture sector was designed in [11]. A plethora of deep learning techniques for crop yield estimation was investigated in [12].

Early grain prediction assists scientists in making better decisions concerning breeding. Utilization of Machine Learning (ML) techniques for integration UAVs and multi-sensor data can enhance the overall crop yield prediction accuracy. Fine tuning weight, via hyper parameter definition using random forest, resulted in the improvement of error rate considerably [13]. Agricultural decision making employing explainable artificial intelligence was investigated in [14]. To ensure early grain prediction by reducing the error rate considerably, a multi-sensor data fusion employing machine learning algorithm was implemented in [15]. To address the issues concerning crop residue burning, a well-enhanced nutrient management method employing integrated technique to concentrate on recycling crop residues can be utilized [16]. To predict agriculture yields in an efficient manner, ML and deep learning were combined in [17]. A case study was investigated in [18] for analyzing crop yield estimation for decision making in agriculture. An accurate mechanism employing IoT devices and ML techniques that can precisely acquire a crop for maximal yield using data of metrological and soil factors were analyzed in [19]. In [20], two different techniques employing data fusion via multimodality and DNN for tea yield estimation was presented.

III. METHODOLOGY

Figure 1 illustrates the workflow of the proposed LZ-GRANN crop yield estimation method using Linear Standardized and Z-score Normalized Class balanced Preprocessing and Gaussian Chebyshev and Radial Artificial Neural Network-based feature selection model. The dataset from [21] is used for training and testing.

As illustrated in Figure 2, first class imbalance and normalization issues are addressed using the Linear Z-score Normalized Class balanced Preprocessing model. The ML algorithm employing Gaussian Chebyshev and Radial Artificial Neural Network is applied to the crop recommendation training model's training dataset. The preprocessed data pertinent features are then selected using ML algorithm to enhance productivity of agricultural practices. Finally, performance parameters are analyzed and validated to interpret the results.

A. Dataset Description

As input to enhance crop yield data, test data obtained from VC Farm Live Data and training data obtained from crop recommendation dataset [21] were utilized. In the recent years smart agriculture through the use of sensors assists the farmers in the decision-making process of their farming strategy. The crop recommendation dataset was constructed by augmenting datasets of rainfall, climate, and fertilizer data available from the Indian Chamber of Food and Agriculture (ICFA).

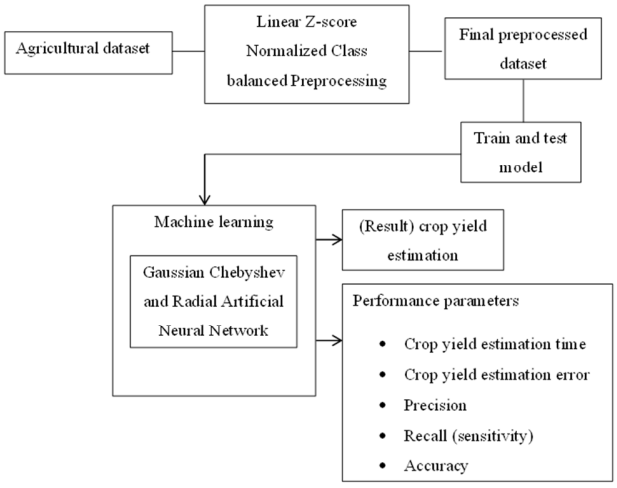

Fig. 1. Structure of Linear Z-score and LZ-GRANN crop yield estimation method.

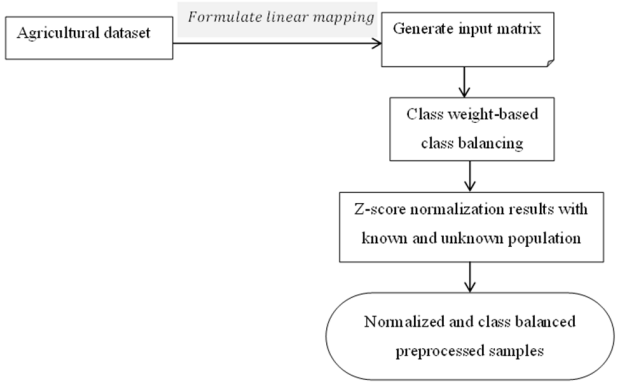

Fig. 2. Structure of Linear Z-score Normalized Class balanced Preprocessing.

B. Linear Z-Score Normalized Class Balanced Preprocessing Model

Preprocessing of raw data can enhance the input quality in providing a compatible composition that controls missing data, recognizes duplicate data, and removes poor data at this level. Additionally, such datasets could include information that is incomplete or imbalance and due to this, such redundant data has to be subjected to filtering and consequently the information has to be normalized. In our work, the Linear Z-score Normalized Class balanced Preprocessing model is employed to obtain preprocessed samples for further processing as seen in Figure 2.

As illustrated in Figure 2 with the agricultural dataset acquired as input, linear maps are employed. Specifically, given two vector spaces S, for samples, and F, for features, over a field D, for dataset, a linear map is formulated as in (1):

$$LT: S \rightarrow F \quad (1)$$

The above linear transformation is harmonious with scalar multiplication and addition as seem in (2):

$$LT(S_i + S_j) = LT(S_i) + LT(S_j) = LT(aS) = aLT(S) \quad (2)$$

Then, with the above set of linear transformations as given in (1) and (2) the input matrix for the raw dataset is formulated as:

$$IM = \begin{bmatrix} S_1F_1 & S_1F_2 & \dots & S_1F_n \\ S_2F_1 & S_2F_2 & \dots & S_2F_n \\ \dots & \dots & \dots & \dots \\ S_mF_1 & S_mF_2 & \dots & S_mF_n \end{bmatrix} \quad (3)$$

The input matrix (3) is formulated from the raw dataset and has to undergo normalization for the corresponding data processing procedure. Initially weights are assigned and to improve class imbalance class they are formulated as seen in (4):

$$W_j = \frac{\sum_{i=1}^{m} s_i}{\sum_{j=1}^{n}(Cl_i * S_{ij})} \quad (4)$$

where the weight for each j class W_j is obtained based on the total numbers of samples s_i in the dataset. Cl_i and S_{ij} represent the total number of unique classes and the total number of rows in the corresponding j class, respectively. With the increase in the training iteration, weights are fine-tuned accordingly. Obtaining the z-score remains the initial step in the normalization process.

The value of Z-score is obtained according to two hypotheses, with and without known mean and standard deviation of population. Population refers to the to the similar items or features which are of interest for experiment, for example, soil pH level sensors, ratio of nitrogen sensors, ratio of phosphorous sensors, etc.

$$z\_K = \frac{S-\mu}{\sigma} \quad (5)$$

From (5), the Z-score of the normalized values z_K is obtained by employing the mean of population μ and standard deviation of population σ respectively. In a similar manner, the Z-score of the normalized values with mean and standard deviation of unknown population is mathematically represented as:

$$z\_UK = \frac{S-S'}{\sigma} \quad (6)$$

From (6), the Z-score normalized values with unknown mean and population z_UK is obtained based on the mean of sample S' and the standard deviation of sample σ respectively. The linear mapping pattern of the normalized samples is:

$$Z\ (LT) = \frac{S-Exp(S)}{\sigma(S)} \quad (7)$$

From (7), a random sample S is normalized by decreasing its expected sample value $Exp(S)$ and dividing the variance by its standard deviation sample values $\sigma(S)$ respectively. Then, the normalized, preprocessed sample, results are formulated as:

$$PS = Z = \frac{S'-Exp(S')}{\sigma(S)/\sqrt{l}}, where S' = \frac{1}{l}\sum_{i=1}^{l} S_i \quad (8)$$

From the results (8), normalized and class-balanced preprocessed samples PS are obtained.

C. Gaussian Chebyshev and Radial Artificial Neural Network-based Feature Selection for Agricultural Big Data Analytics

To implement the necessary measures to enhance crop yield, selecting relevant features is pivotal to achieve sufficient precision and accuracy. The raw feature in hand may contain irrelevant and redundant attributes and these attributes require to be discarded. In this section, optimal feature selection for enhancing crop yield estimation using Gaussian Chebyshev and Radial Artificial Neural Network-based feature selection model is designed.

The Gaussian Chebyshev and Radial Artificial Neural Network-based feature selection model, as presented in Figure 3, consists of a set of neurons, arranged in a connected network comprising of three layers: input layer, hidden layer, and output layer. The input layer is an input vector with each element denoting a feature, denoted as F, hidden layer includes Radial Basis Neurons (RBN) that generate hidden patterns using Gaussian basis function and Chebyshev distance therefore forming the basis of big data analytics. By uncovering the hidden patterns prominent features can be selected. Finally, the output layer includes Linear Neurons.

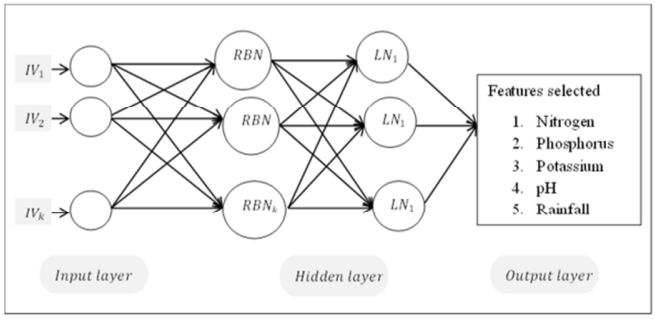

Fig. 3. Structure of Gaussian Chebyshev and Radial Artificial Neural Network-based feature selection.

In Figure 3, a set of neurons or preprocessed samples PS are linked together to form a ternary layer network. All neurons in the hidden layer are linked from the subsequent input layer and then to every neuron in the output layer. Furthermore, a link exists between two adjoining layers. Initially, an external input of PS is fed into each of k-neurons in the hidden layer. The output, or optimal features, obtained by each hidden neuron, is fed into all the neurons of the output layer. To start with the input vector formalizing the input layer figure, IV represents the d dimensional input vector as seen in (9).

$$IV = (PS_1, PS_2, \dots, PS_d)^T U (F_1, F_2, \dots, F_d)^T \quad (9)$$

From (9), the input vector in the input layer is formulated taking into consideration the preprocessed samples $(PS_1, PS_2, \dots, PS_d)^T$ and features $(F_1, F_2, \dots, F_d)^T$ as input. Secondly, the hidden patterns are to be interpreted to generate agri-intelligence. RBN is employed using Gaussian basis function and Chebyshev distance to interpret the hidden

patterns in the generated results to obtain the features according to the significance level. Let $C_i(t) = [C_{1i}(t), C_{2i}(t), \ldots\ldots, C_{ni}(t)]^T$ denote the center vector, $|PS - C|$ denote the distance between PS and C and therefore form a Gaussian basis function:

$$\varphi(IV, C, \alpha) = Exp\left[-\frac{(IV-C)^2}{2\alpha^2}\right] \quad (10)$$

From (10), $IV - C$ forms the Chebyshev distance expressed as:

$$Dis_{Chebyshev}(IV, C) = \max_i(|IV - C|) \quad (11)$$

$$Dis_{Chebyshev}(IV, C) = \max(|C_2 - C_1|, |IV_2 - IV_1|) \quad (12)$$

From (11) and (12), the perfect subsets are explored by validating the perfectly generated subsets. By interpreting the hidden patterns in the generated results with the objective of avoiding local optima and to explore broad search space to arrive at a selected set of features and avoid early convergence, Linear Function is applied at the output layer.

$$Out_j = \sum_{i=1}^{k} W_{ji} Exp\left\{-\frac{\left\|IV' - C_i'\right\|^2}{2\alpha^2}\right\} + W_{jo} \quad (13)$$

From (13), the features selected in the output layer W_{ji} are obtained by considering association weight connecting the *i*th hidden neuron to the *j*th output neuron, W_{jo} is the bias connecting the *j* output neuron, the hidden neuron center vector C_i and the activation function α^2. The results obtained from Out_j determine the informative features.

IV. EXPERIMENTAL SETUP

In this section, performance evaluation of the proposed method is analyzed and validated in Python with the aid of the employed agricultural dataset [21]. Simulations are analyzed using five performance parameters, namely crop yield estimation time, crop yield estimation error, precision, recall, and accuracy. To ensure fair comparison, similar samples are employed from the training and testing dataset for all the three considered methods, LZ-GRANN, MMLA and GAN-CNN. The training-testing ratio was 80:20. Ten-fold cross-validation was utilized for measuring the results.

A. Performance Analysis of Crop Yield Estimation Time

Crop yield estimation time is a paramount performance parameter. Table I lists the values of the evaluation measures and the crop yield estimation time for LZ-GRANN, MMLA and GAN and CNN different crop yield estimation methods.

The crop yield estimation time calculated with the proposed LZ-GRANN method was observed to be 17.5 ms for 50 samples, 25.35 ms for 100 samples, and 31.55 ms for 150 samples. The crop yield estimation time acquired with the existing MMLA method was 23.5 ms for 50 samples, 35.55 ms for 100 samples, and 41.35 ms for 150 samples. The crop yield estimation time acquired with GAN and CNN was 27.5 ms for 50 samples, 45.35 ms for 100 samples, and 55.35 ms for 150 samples. It can be concluded that the proposed LZ-GRANN consumed lesser crop yield estimation time than the existing methods.

TABLE I. CROP YIELD ESTIMATION TIME OF LZ-GRANN METHOD WITH MMLA, GAN AND CNN

Samples	Crop yield estimation time (ms)		
	LZ-GRANN	MMLA	GAN and CNN
50	17.5	23.5	27.5
100	25.35	35.55	45.35
150	31.55	41.35	55.35
200	40	45.25	68.35
250	43.85	50.35	75.35
300	35.35	42.45	60.25
350	30	38.35	55.25
400	28.55	35.15	48.35
450	35.35	43.25	51.35
500	41.35	50.55	55.25

B. Performance Analysis of Crop Yield Estimation Error Rate

In this section, the crop yield estimation error rate or number of misclassifications are measured. Table II lists the values of the evaluation measures and the crop yield estimation error for LZ-GRANN, MMLA and GAN, and CNN. The crop yield estimation error of the proposed LZ-GRANN method was 8% for 50 samples, 10% for 100 samples, and 14% for 150 samples. MMLA crop yield estimation error was 14% for 50 samples, 16% for 100 samples, and 19% for 150 samples. GAN and CNN result was 18% for 50 samples, 21% for 100 samples and 24% for 150 samples. The proposed model of LZ-GRANN incurred lesser crop yield estimation error than the existing methods.

TABLE II. CROP YIELD ESTIMATION ERROR OF LZ-GRANN METHOD WITH MMLA, GAN, AND CNN

Samples	Crop yield estimation error (%)		
	LZ-GRANN	MMLA	GAN and CNN
50	8	14	18
100	10	16	21
150	14	19	24
200	16	21	25
250	18	19	23
300	14	17	20
350	13	15	18
400	11	13	16
450	9	11	14
500	12	14	17

C. Performance Analysis of Precision, Recall, and Accuracy

In this section, the precision, recall and accuracy involved in the crop yield estimation for interpreting the hidden patterns to enhance the crop yield is analyzed and validated. Table III lists the acquired values of the evaluation measures of the considered crop yield estimation methods.

TABLE III. PRECISION, RECALL, AND ACCURACY COMPARISON

	LZ-GRANN	MMLA	GAN and CNN
Precision	0.93	0.86	0.81
Accuracy	0.92	0.85	0.78
Recall	0.97	0.96	0.94

For the given crop recommendation dataset, the existing methods MMLA and GAN and CNN have lower precision, accuracy, and recall values than the proposed LZ-GRANN

method. As a conclusion, the LZ-GRANN method improves overall crop yield estimation by selecting more accurately features.

V. CONCLUSION

Linear Z-score and Gaussian Radial Artificial Neural Network-based (LZ-GRANN) constitutes a crop yield estimation method that is designed with minimum estimation time and error rate to enhance productivity of agricultural practices. Initially, the input vector matrix is formulated with the raw dataset used as input. On the next step, the ideal Z-score normalized values with known and unknown mean and population are chosen and applied to obtain normalized samples with a linear mapping pattern. Furthermore, Gaussian Chebyshev and Radial Artificial Neural Network-based feature selection is applied to the preprocessed samples as input in the input layer, generating hidden patterns via Gaussian Chebyshev and Chebyshev distance to traverse extensive search space. Finally, samples and selected features are given as input and the linear function is applied at the output layer to generate the results in the output matrix.

Compared to the existing models Multi-sensor Machine Learning Approach (MMLA), Generative Adversarial Network (GAN), and Convolutional Neural Network (CNN), the proposed LZ-GRANN consumed less crop yield estimation time, incurred less crop yield estimation error, and improved overall crop yield estimation in terms of precision, accuracy and recall.

REFERENCES

[1] 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," *IEEE Access*, vol. 11, pp. 20795–20805, Feb. 2023, https://doi.org/10.1109/ACCESS.2023.3249205.

[2] J. Zhang, H. Tian, P. Wang, K. Tansey, S. Zhang, and H. Li, "Improving wheat yield estimates using data augmentation models and remotely sensed biophysical indices within deep neural networks in the Guanzhong Plain, PR China," *Computers and Electronics in Agriculture*, vol. 192, Jan. 2022, https://doi.org/10.1016/j.compag.2021.106616.

[3] M. Van der Velde and L. Nisini, "Performance of the MARS-crop yield forecasting system for the European Union: Assessing accuracy, in-season, and year-to-year improvements from 1993 to 2015," *Agricultural Systems*, vol. 168, pp. 203–212, Jan. 2019, https://doi.org/10.1016/j.agsy.2018.06.009.

[4] D. Paudel *et al.*, "Machine learning for large-scale crop yield forecasting," *Agricultural Systems*, vol. 187, Dec. 2020, https://doi.org/10.1016/j.agsy.2020.103016.

[5] R. Priyatikanto, Y. Lu, J. Dash, and J. Sheffield, "Improving Generalisability and Transferability of Machine-Learning-Based Maize Yield Prediction Model Through Domain Adaptation," May 28, 2022, *Rochester, NY*: 4122021, https://doi.org/10.2139/ssrn.4122021.

[6] Y. Wang, W. Shi, and T. Wen, "Prediction of winter wheat yield and dry matter in North China Plain using machine learning algorithms for optimal water and nitrogen application - ScienceDirect," *Agricultural Water Management*, vol. 277, Mar. 2023, https://doi.org/10.1016/j.agwat.2023.108140.

[7] H. Tian *et al.*, "A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the Guanzhong Plain, PR China," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, https://doi.org/10.1016/j.jag.2021.102375.

[8] T. Iizumi, Y. Shin, W. Kim, M. Kim, and J. Choi, "Global crop yield forecasting using seasonal climate information from a multi-model ensemble," *Climate Services*, vol. 11, Jul. 2018, https://doi.org/10.1016/j.cliser.2018.06.003.

[9] A. Mateo-Sanchis, M. Piles, J. Muñoz-Marí, J. E. Adsuara, A. Pérez-Suay, and G. Camps-Valls, "Synergistic integration of optical and microwave satellite data for crop yield estimation," *Remote Sensing of Environment*, vol. 234, Dec. 2019, https://doi.org/10.1016/j.rse.2019.111460.

[10] M. Rashid, B. Bari, Y. Yusup, M. Kamaruddin, and N. Khan, "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction," *IEEE Access*, vol. 9, Apr. 2021, https://doi.org/10.1109/ACCESS.2021.3075159.

[11] E. Elbaşı *et al.*, "Artificial Intelligence Technology in the Agricultural Sector A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 171–202, Jan. 2023.

[12] A. Oikonomidis, C. Catal, and A. Kassahun, "Deep learning for crop yield prediction: a systematic literature review," *New Zealand Journal of Crop and Horticultural Science*, vol. 51, pp. 1–26, Feb. 2022, https://doi.org/10.1080/01140671.2022.2032213.

[13] Padma T. and D. Sinha, "Crop Yield Prediction Using Improved Random Forest," *ITM Web Conference*, vol. 56, no. 02007, Aug. 2023, https://doi.org/10.1051/itmconf/20235602007.

[14] M. Shams, S. Adel Gamel, and F. M. Talaat, "Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making," *Neural Computing and Applications*, vol. 36, Jan. 2024, https://doi.org/10.1007/s00521-023-09391-2.

[15] S. Fei *et al.*, "UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat," *Precision Agric*, vol. 24, no. 1, pp. 187–212, Feb. 2023, https://doi.org/10.1007/s11119-022-09938-8.

[16] A. K. Bhardwaj *et al.*, "Residue recycling options and their implications for sustainable nitrogen management in rice–wheat agroecosystems," *Ecological Processes*, vol. 12, no. 1, Nov. 2023, Art. no. 53, https://doi.org/10.1186/s13717-023-00464-7.

[17] P. Sharma, P. Dadheech, N. Aneja, and S. Aneja, "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning," *IEEE Access*, vol. 99, pp. 1–1, Jan. 2023, https://doi.org/10.1109/ACCESS.2023.3321861.

[18] G. Sahbeni, B. Székely, P. K. Musyimi, G. Timár, and R. Sahajpal, "AgriEngineering | Free Full-Text | Crop Yield Estimation Using Sentinel-3 SLSTR, Soil Data, and Topographic Features Combined with Machine Learning Modeling: A Case Study of Nepal," *AgriEngineering*, vol. 5, pp. 1766–1788, Oct. 2023, https://doi.org/10.3390/agriengineering5040109.

[19] A. Ikram *et al.*, "Crop Yield Maximization Using an IoT-Based Smart Decision," *Journal of Sensors*, vol. 2022, no. 1, pp. 1–15, May 2022, https://doi.org/10.1155/2022/2022923.

[20] Z. Ramzan, H. Asif, I. Yousuf, and M. Shahbaz, "A Multimodal Data Fusion and Deep Neural Networks Based Technique for Tea Yield Estimation in Pakistan Using Satellite Imagery," *IEEE Access*, vol. 99, pp. 1–1, Jan. 2023, https://doi.org/10.1109/ACCESS.2023.3271410.

[21] N. Gaud, "Crop Recommendation System using Machine Learning," 2022, [Comma Separated Values], Available: https://www.kaggle.com/code/nirmalgaud/crop-recommendation-system-using-machine-learning/input.