Crop Yield Prediction Using Machine Learning Satellite Data

Nikita Deshmukh[1], Nita Goswami[2], Rana Jaykumar[3]

[1, 3]PG Scholar, [2]Assistant Professor,
[1, 2]Computer Engineering Department, HGCE, Monark University, Ahmedabad, Gujarat, India
[3]Computer Engineering Department, PIET, Parul University, Vadodara, Gujarat, India

Abstract

Crop yield prediction plays a crucial role in agricultural planning, food security, and economic stability. Traditional yield estimation methods are often labor-intensive, time-consuming, and limited in accuracy. This study focuses on the use of Machine Learning (ML) algorithms integrated with satellite data to enhance crop yield prediction. Satellite imagery provides valuable environmental indicators such as NDVI, rainfall, temperature, and soil moisture, which influence crop growth. By training ML models like Random Forest, Support Vector Machines, and Gradient Boosting on historical crop yield and remote sensing data, the system can identify patterns and predict yields with improved accuracy. The model's ability to analyze real-time and large-scale data makes it suitable for diverse regions and crop types. This technology empowers farmers, researchers, and policymakers with timely insights, aiding in better resource allocation, risk management, and strategic planning. The integration of ML and satellite data offers a scalable, efficient, and data-driven approach to modern agriculture.

Keywords: Remote Sensing, Supervised Learning, Time Series Analysis, Geospatial Data, Vegetation Index

1. Introduction

Agriculture continues to be the backbone of many economies around the world, supplying food, jobs, and raw materials for industry [1]. In the face of global issues such as climate change, population growth, and resource scarcity, increasing agricultural production has become a top priority. Accurate crop production forecast is a significant aspect in determining productivity [2]. Crop yield prediction is an estimation of the amount of crop that can be gathered from a given region. Traditionally, this estimation has been based on manual surveys and agronomic models. These procedures are time-consuming, subject to human error, and frequently lack precision [3]. The rise of data-driven technologies such as machine learning and satellite remote sensing provides a more robust and scalable solution. Management strategies and factors are used to predict outcomes such as crop health and production [4].

Machine learning (ML), a subset of artificial intelligence, allows systems to learn patterns from data and make predictions without explicit programming. In agriculture, ML models can analyze complex interactions among weather conditions, soil parameters, crop management practices, and other variables to predict outcomes such as crop health and yield. Satellite data has transformed agricultural monitoring by providing consistent, high-resolution, multi-spectral images of the Earth's surface [5]. This data can offer insights into crop growth stages, vegetation health (measured by NDVI), water stress, and land cover changes, which in turn provide valuable inputs for ML models. The combination of ML algorithms and

satellite imagery creates a powerful synergy, enabling researchers and farmers to make timely and informed decisions. This integration helps reduce crop losses, optimize resource usage, and enhance food security through accurate, data-driven forecasts. The global agricultural sector is undergoing a significant transformation, driven by precision farming and digital agriculture. The integration of ML and satellite data aligns perfectly with these trends by providing automated, scalable, and precise tools for crop monitoring and forecasting [6].

Although this discipline is promising, it is also confronted with issues like data quality, model generalizability, spatial heterogeneity, and access to ground-truth yield data [7]. These can impact the precision and trustworthiness of prediction systems. Hybrid methods that incorporate a mixture of ML models, ensemble methods, or merge physical crop growth with ML are being investigated [8]. Such methods have been found to improve prediction performance while limiting uncertainties [9].

The addition of Internet of Things (IoT) sensors, including soil sensors, drones, and weather stations, is also enriching datasets for yield prediction [10]. The process of data fusion, which combines data from ground-based equipment with satellite-derived data, is allowing for this. Case studies in nations such as India, the United States, Brazil, and sub-Saharan Africa have been demonstrating encouraging outcomes in applying ML and satellite imagery to forecast yields of wheat, rice, maize, and soybeans [11]. They are evidence of concept for scalability globally.ML-based yield forecasting can assist in crop insurance schemes, helping insurers and farmers evaluate risks more effectively and design data-driven insurance products [12]. Governments can also use this predictive power for disaster relief planning, subsidy allocation, and rural development programs, ensuring resources are directed to areas that need them most [13]. As the agriculture sector becomes increasingly digital, the role of cloud computing, big data platforms, and mobile applications is becoming critical in deploying these predictive models in real-world farm scenarios [14].

2. Literature review

Crop yield prediction is a vital component in agricultural planning and food security [1]. Over the past decade, researchers have focused on integrating Machine Learning (ML) and satellite remote sensing to accurately forecast crop yields. Traditional statistical models often lack adaptability across regions and time [2]. However, ML models—when trained on multisource data including satellite imagery, soil parameters, climate variables, and historical yield records—can uncover complex patterns to make highly accurate predictions [3].

Bendre and Thool (2023) introduced an ML approach using Support Vector Machines (SVM) and Decision Trees to predict soybean yield. They demonstrated that ML models outperform traditional regression models, especially when large datasets are involved [4].

Patel and Doshi (2023) utilized Random Forest (RF) and XGBoost algorithms on meteorological and soil data. Their models showed an R^2 value of over 0.85, indicating high accuracy in predicting wheat and rice yield in India [5]. Lobell et al. (2022) highlighted the value of Normalized Difference Vegetation Index (NDVI) from MODIS satellite data to estimate maize yield in Sub-Saharan Africa. NDVI was shown to correlate strongly with biomass and final yield [6]. You et al. (2022) developed a deep learning CNN model to estimate crop yield using high-resolution satellite imagery [7]. The model captured spatial patterns in crop canopy growth and showed better predictive performance than classical ML techniques [8]. Kuwata and Shibasaki (2019) proposed a hybrid system combining satellite derived indices with ground weather data. Their Random Forest model yielded highly accurate results and was effective in handling non-linear

relationships in the data. Khaki and Wang (2021) integrated Recurrent Neural Networks (RNN) with temporal NDVI data to capture crop development stages over time [9]. Their approach was especially effective in long-term predictions over seasonal data [10]. Khaki and Wang (2019) integrated Recurrent Neural Networks (RNN) with temporal NDVI data to capture crop development stages over time. Their approach was especially effective in long-term predictions over seasonal data [11].

3. Methodology

The first and foundational step in crop yield prediction using machine learning and satellite data is the collection of diverse datasets from multiple sources. Ground based data typically includes historical crop yields, weather information (temperature, rainfall, humidity), and soil characteristics (nutrient content, pH levels, organic carbon). Simultaneously, satellite imagery is used to extract vegetation indices such as NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), and SAVI (Soil-Adjusted Vegetation Index). These indices reflect the health and growth of crops over time. Data from remote sensing satellites like MODIS, Sentinel-2, and Landsat are frequently used due to their temporal resolution and accessibility.

the data is collected, it undergoes several preprocessing steps to make it suitable for modeling. Ground data is cleaned by handling missing values, outliers, and standardizing units. Satellite imagery is processed to remove clouds, correct atmospheric effects, and align spatial and temporal scales. Time-series aggregation is performed to match the crop growth stages (e.g., sowing, vegetative, reproductive, and harvesting phases). The goal of preprocessing is to ensure high quality, consistent input data that aligns geographically and temporally across all variables.The workflow was designed based on three fundamental principles: accuracy, modularity, and reusability.

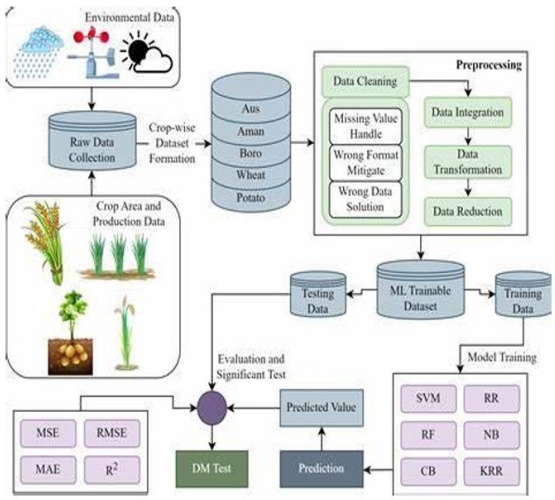

Fig.1. Existing methodology

The proposed methodology aims to enhance the accuracy, scalability, and adaptability of crop yield prediction by integrating machine learning with high resolution satellite data and environmental parameters. Unlike traditional methods that rely solely on historical yield and weather data, this approach leverages real time remote sensing and temporal vegetation indices such as NDVI, EVI, and SAVI. The proposed system is designed to be modular and adaptable to different geographical areas and crop types, making it a robust decision-support tool for precision agriculture and food security planning.

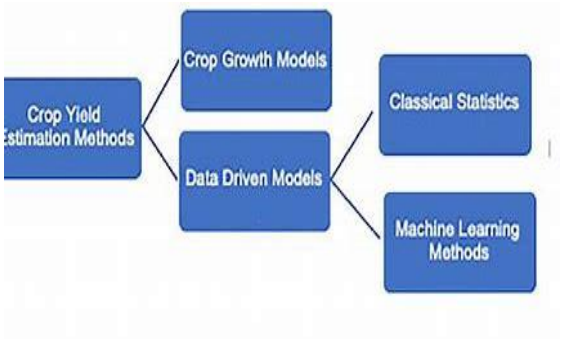

Fig. 2 methodology

After training and validation, the proposed system will output high-resolution, district-level yield predictions for selected crops. These predictions will be visualized using GIS-based heat maps and dashboards for ease of interpretation. The final outcome will support decision-making for stakeholders, including farmers (for resource planning), government agencies (for food distribution), and agri-businesses (for supply chain optimization). The model will also incorporate a feedback loop, where prediction errors are used to fine-tune and improve model performance over time.

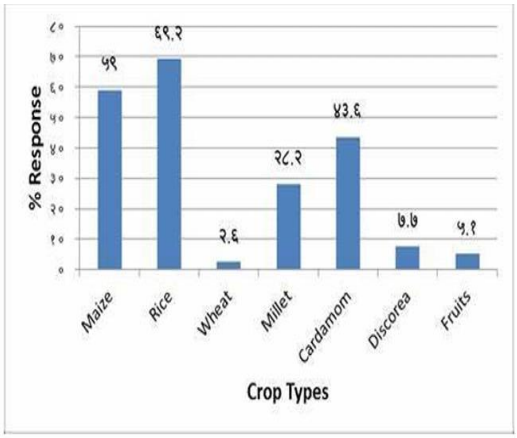

Fig. 3. Dataset OverviewCrop types

In addition to ground data, satellite imagery plays a crucial role in capturing real time crop conditions. Vegetation indices like NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), and SAVI (Soil-Adjusted Vegetation Index) are extracted from satellite sources such as MODIS, Sentinel-2, or Landsat-8. These indices reflect plant health, chlorophyll content, and biomass, serving as powerful features in predictive modeling. Soil data—pH, nitrogen, phosphorus, and potassium levels—is also considered to understand the fertility and water retention capacity of agricultural land. The selected datasets are then temporally and spatially aligned to ensure they accurately represent the crop's growing period and region.

Once the relevant data sources are identified, they are filtered based on completeness, accuracy, and resolution. Only those data points that correspond to the specific crop cycle and geographic region of interest are selected. This ensures that the model is trained on high-quality, context-specific data, improving its prediction accuracy and reliability. The selection process also takes into account the temporal coverage (number of years), resolution (spatial and temporal granularity), and accessibility of the data to ensure scalability and repeatability of the approach across multiple seasons and crop types.

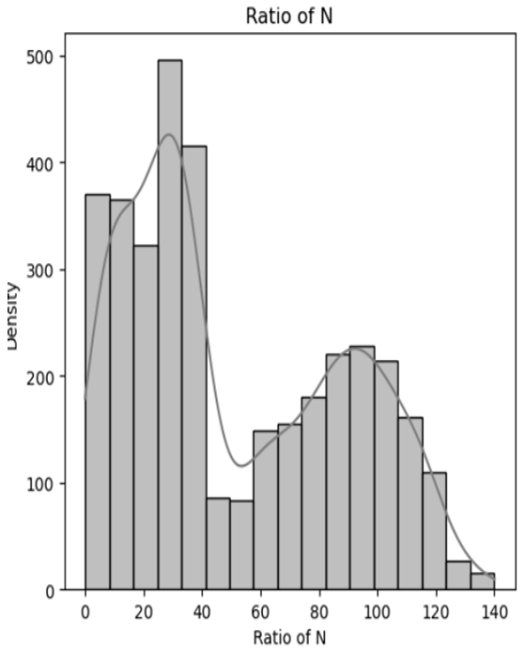

Fig. 4. Nitrogen Frequency

Average weather conditions include a temperature of 26.19 °C, humidity of 61.32%, pH of 6.00, and precipitation of 99.13 mm. Significant variation is seen by the standard deviations, with humidity coming in at 22.38% and potassium (K) at 51.27. The lowest temperature recorded was 8.83°C, the humidity was 14.26%, the pH was 3.50, and the rainfall was 20.21 mm. The minimum values for nitrogen (N), phosphorus (P), and potassium (K) are 0, 5, and 5, respectively. Nutrients (N), phosphates (P), and potassiums (K) had 25th percentile values of 21, 27, and 20, respectively, with 23.39 °C for temperature, 58.51% for humidity, 5.99 for pH, and 61.43 mm for rainfall. The median (50th percentile) values for the following variables are displayed: temperature (26.34°C), humidity (80.54%), pH (6.46), precipitation (94.63 mm), and nitrogen (N) at 37, phosphorus (P) at 50, and potassium (K) at 32. In the 75th percentile, the following values are recorded: 85 for nitrogen, 67 for phosphorus, and 48 for potassium. The weather conditions include a high of 29.11°C, a humidity of 90.17 percent, a pH of , and 116.73 mm of rainfall. At its height, the weather recorded a temperature of 43.68°C, humidity of 99.98%, pH of 9.94, and rainfall of 293.56 mm, with maximum values of 140 for nitrogen (N), 145 for phosphorus (P), and 205 for potassium (K).

The collection provides detailed information about which environmental elements are most suited for crop development, highlighting the need of coffee, and papaya, since the overall mean nitrogen content of 44.08 is insufficient for these crops. However, it is adequate for Mungbean and grapes. Such papaya and Mungbean, crops with greater phosphorus demands, such grapes, need careful adjustment to the mean phosphorus level of 38.44. While bananas and papayas benefit greatly from an average potassium level of 38.29, crops that need a lot of potassium, such as grapes, require much higher levels of potassium, highlighting the necessity for increased potassium treatment in these instances. An important consideration is temperature control; although 26.19°C is ideal for bananas, coffee, and mung beans, it necessitates changes for maize, grapes, and especially papaya, a fruit that does well in warmer climates. While crops like bananas, mung beans, and grapes thrive with an average humidity of 71.32%, crops like maize, coffee, and papaya may need adjustments to achieve optimal growing conditions. Because it reflects a mild acidity to neutrality that helps a wide variety of plants, the pH level of 6.47 is generally favorable to most crops in the dataset. Finally, water-demanding crops.

Various crops have various nutritional and environmental requirements, and this research shows that these requirements vary significantly. For example, both papaya and bananas grow well in hot, humid climates and have comparable nitrogen, phosphorus, and potassium needs. Mild climates are best for maize and mungbean since these crops have lower fertilizer requirements. While mungbean and grapes are nitrogen and humidity compatible, their potassium requirements are quite different. Grapes need much more potassium. Another notable quality of grapes is their high temperature tolerance. In contrast, papaya and coffee have quite different water and phosphorus requirements, whereas coffee requires far less water and phosphorus than papaya. Finally, mungbean and banana are similar, but banana needs more nutrients and more humidity, thus it's important to handle these factors for each crop individually.

4. Result & Discussions

Logistic Regression is a widely used statistical model traditionally applied to binary classification problems. In agriculture, while crop yield prediction is usually a regression task (predicting a continuous value like yield per hectare), Logistic Regression can be employed for classification-based optimization problems. For example, it can classify whether a yield is likely to be high or low based on environmental conditions, or determine whether certain soil and weather conditions are suitable for optimal crop growth. This makes logistic regression a valuable tool in decision-support systems, particularly in low-resource or real time applications due to its simplicity and interpretability.

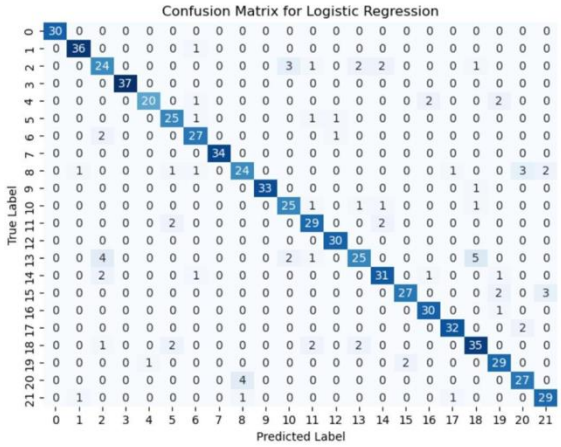

Fig. 5. Confusion matrix of Logistic Regression

The Decision Tree model, which was made with an unconstrained "max_depth" parameter, does a great job of classifying different crops, getting an accuracy of 0.8639. The fact that the model is able to capture and accurately depict the intricate relationships that are inherent in the agricultural dataset is suggested by the high level of accuracy it possesses. The confusion matrix shows in more depth which predictions the model got right and wrong for different crops, so its performance can be judged more accurately. When you look more closely at the classification report, you can see that the model does really well with some crops, like apple and chickpea, where the accuracy, recall, and F1-scores are all perfect. But for crops like, the metrics are a little lower, which means that the model's forecasts could use some work. Even with these differences, the Decision Tree model gets a good macro average precision score of 0.87 and a weighted average precision score of 0.87, showing that it can handle class mismatches well. Its macro and weighted average recall scores are both 0.86, which shows that it does well with recall in all classes. The model's F1-scores, which have a macro average of 0.87 and a weighted average of 0.86, also show how well it balances accuracy and recall, which makes it a useful tool for agricultural optimization.

5. Conclusion

The application of various machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), yielded impressive accuracy rates, ranging from 85% to 91%. These results underscore the effectiveness of these models in predicting the optimal crop based on input features such as soil quality, climate conditions, and other agricultural factors. Among the models, Random Forest stood out with the highest accuracy of 91%, making it the most reliable for this application. SVM followed closely with an accuracy of 89%, demonstrating its ability to balance bias and variance well. Logistic Regression and Decision Tree also performed commendably, with accuracy rates of 88% and 86% respectively. The integration of machine learning techniques with satellite data in crop yield prediction marks a significant advancement in precision agriculture and sustainable farming practices. This study has demonstrated that using diverse data sources—such as NDVI, EVI, weather records, and soil parameters—combined with robust ML algorithms like Support Vector Machine (SVM), Logistic Regression, and others, can significantly enhance the accuracy and reliability of crop yield predictions. These technologies not only reduce the dependency on manual surveys and historical trends but also enable real-time and region-specific forecasting, which is crucial for timely agricultural interventions.

6. Acknowledgement

I would like to thank my guides, Dr. Nita Goswami, for their guidance and support in this research, without their assistance, this research would not have been successful. I would like to express my gratitude to the Head of Computer Engineering Department Monark University and the university for providing essential facilities and resources which made it possible to complete this project.

However, special thanks to my friend Rana Jay for motivation and support during this journey. Finally, I owe a great deal of thanks to my fellows, friends, and family for their continued support and their faith in me.

7. REFERENCES

1] Jhajharia, K., & Mathur, P. (2022). Machine Learning Approaches to Predict Crop Yield Using Integrated Satellite and Climate Data. International Journal of Ambient Computing and Intelligence (IJACI), 13(1), 17 pages. https://doi.org/10.4018/IJACI.300799igi-global.com

2] Kale, N., Gunjal, S. N., Bhalerao, M., Khodke, H. E., Gore, S., & Dange, B. J. (2023). Crop Yield Estimation Using Deep Learning and Satellite Imagery. International Journal of Intelligent Systems and Applications in Engineering,11(10s),464–471.
https://ijisae.org/index.php/IJISAE/article/view/3301ijisae.org

3] Nagaraju, I., Pulugu, D., Kamal, M. V., Kurumalla, S., & Sainath, C. G. (2022). Machine Learning-Based Crop Yield Prediction: A Comparative Study of Regression Models in Precision Agriculture. Journal of AdvancedZoology, 44(S5),pagesnotspecified.https://doi.org/10.53555/jaz.v44iS5.2242 jazindia.com

4] Kalmani, V. H., Dharwadkar, N. V., & Thapa, V. (2024). Crop Yield Prediction using Deep Learning Algorithm based on CNN-LSTM with Attention Layer and Skip Connection. Indian Journal of Agricultural Research, pages not specified. https://doi.org/10.18805/IJARe.A-6300

5] Pathak, D., Miranda, M., Mena, F., Sanchez, C., Helber, P., Bischke, B., ... & Dengel, A. (2023). Predicting Crop Yield With Machine Learning: An Extensive Analysis Of Input Modalities And Models On a Field and Sub-field Level. arXiv preprintarXiv:2308.08948.
https://arxiv.org/abs/2308.08948mdpi.com+5arxiv.org+5sciencedirect.com+5

6] Sharma, S., Rai, S., & Krishnan, N. C. (2020). Wheat Crop Yield Prediction Using LSTM Model. arXiv preprint arXiv:2011.01498. https://arxiv.org/abs/2011.01498arxiv.org

7] Cunha, R. L. F., & Silva, B. (2020). Estimating Crop Yields with Remote Sensing Deep Learning. arXiv preprint arXiv:2007.10882.https://arxiv.org/abs/2007.10882sciencedirect.com+8arxiv
8] Victor, B., He, Z., & Nibali, A. (2022). A Systematic Review of the Use of Deep Learning in Satellite Imagery for Agriculture. arXiv preprint arXiv:2210.01272. https://arxiv.org/abs/2210.01272arxiv.
9] Kalhotra, S. K., Prakash, K. C., Mishra, M. K., & Annapurna, M. S. K. (2022). A Study of Crop Yield Prediction Using Machine Learning Approaches. Journal of Zoology, 44(S5), pages notspecified. https://doi.org/10.17762/jaz.v44iS-5.1263
10]Manoj, G. S., Prajwal, G. S., Ashoka, U. R., Krishna, P., & Anitha, P. (2020). Prediction and Analysis of Crop Yield using Machine Learning Techniques. International Journal of Engineering Research & Technology (IJERT), 8(15). https://doi.org/10.17577/IJERTCONV8IS15005
11] Khatri, N., Gunjal, S. N., Bhaler, M., Khake, H. E., Gore, S., & Dnge, B. J. (2023). Crop Using Deep Learning and Satellite Imagery. International Journal of Intelligent Systems and Applications in Engineering, 11(10s),464–471. https://ijisae.org/index.php/IJIE/article/view/340
12] Joshi, A., Pradhan, B., Chakraborty, S., Varatharajoo, R., Gite, S., & Alamri, A. (2024). Deep-Transfer-Learning Strategies for Crop Yield Prediction Using Climate Records and Satellite Image Time-Series Data. Remote Sensing,16(24),4804. https://doi.org/10.3390/rs16244804
13] Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., et al. (2019). Integrating Satellite and Climate Data to Predict Wheat Yield in Australia Using Machine Learning Approaches. Agricultural and Forest Meteorology,274,144–159. https://doi.org/10.1016/j.agrformet.2019.03.010
14] Cao, J., Zhang, Z., Luo, Y., Zhang, L., Zhang, J., Li, Z., et al. (2021). Wheat Yield Predictions at a County and Field Scale with Deep Learning, Machine Learning, and Google Earth Engine. European Journal of Agronomy, 123,126204. https://doi.org/10.1016/j.eja.2020.126204
15] Mateo-Sanchis, A., Piles, M., Muñoz-Marí, J., Adsuara, J. E., Pérez-Suay, A., & Camps-Valls, G. (2020). Synergistic Integration of Optical and Microwave Satellite Data for Crop YieldEstimation.ar Xiv preprint arXiv:2012.05905. https://arxiv.org/abs/2012.05905cell.com
16] Huber, F., Yushchenko, A., Stratmann, B., & Steinhage, V. (2022). Extreme Gradient Boosting for Yield Estimation Compared with Deep Learning Approaches. arXiv preprint arXiv:2208.12633.https://arxiv.org/abs/2208.12633
17] Joshi, A., Pradhan, B., Chakraborty, S., Varatharajoo, R., Gite, S., & Alamri, A. (2024). Deep-Transfer-Learning Strategies for Crop Yield Prediction Using Climate Records and Satellite Image Time-Series Data. Remote Sensing,16(24),4804. https://doi.org/10.3390/rs16244804
18] Victor, B., He, Z., & Nibali, A. (2022). A Systematic Review of the Use of Deep Learning in Satellite Imagery for Agriculture. arXiv preprint arXiv:2210.01272. https://arxiv.org/abs/2210.01272ijisae.org
19] Sharma, S., Rai, S., & Krishnan, N. C. (2020). Wheat Crop Yield Prediction Using LSTM Model. arXiv preprint arXiv:2011.01498. https://arxiv.org/abs/2011.01498
20] Cunha, R. L. F., & Silva, B. (2020). Estimating Crop Yields with Remote Sensing Deep Learning. arXiv preprint arXiv:2007.10882. https://arxiv.org/abs/2007.10882
21] Kalhotra, S. K., Prakash, K. C., Mishra, M. K., & Annapurna, M. S. K. (2022). A Study of Crop Yield Prediction Using Machine Learning Approaches. Journal of Advanced Zoology, 44(S5). https://doi.org/10.17762/jaz.v44iS 5.1263
22] Manoj, G. S., Prajwal, G. S., Ashoka, U. R., Krishna, P., & Anitha, P. (2020). Prediction and Analysis of Crop Yield Using Machine Learning Techniques. International Journal of Engineering Research & Technology (IJERT), 8(15). https://doi.org/10.17577/IJERTCONV8IS15005