International Journal of
ADVANCED RESEARCH & INNOVATIONS

Volume 15, Issue-8
August, 2025, P04
ISSN: 2319-9253

A DATA-DRIVEN APPROACH TO CROP YIELD PREDICTION USING ADVANCED MACHINE LEARNING TECHNIQUES

[1]Dr.T.Veeranna, [2]B.Yashwanth, [3]J.Sivanagaraju, [4]S.Chaitanya Reddy,
[5]D.Reshma Priya, [6]J.Praveen
[1]Associate Professor, Dept. of CSE (AI&ML), Sai Spurthi Institute of Technology, Khammam, Telangana, India.
[2,3,4,5,6]B.Tech Student, Dept. of CSE(AI&ML), Sai Spurthi Institute of Technology, Khammam,Telangana,India.

Abstract: Half or more of India's population relies on agriculture for their livelihood, making it an essential sector of the Indian economy. The future of agriculture is in jeopardy due to the growing threat posed by climate change and other environmental factors. Improving decision-making about agricultural cultivation and growing practices, machine learning (ML) offers a tool for crop yield prediction (CYP). Several approaches have been devised to analyze AI-based crop yield prediction algorithms; this study centers on a systematic review that extracts and synthesizes CYP features. Less relative inaccuracy and less capacity to predict crop yield are the primary drawbacks of neural networks. Supervised learning algorithms had a hard time selecting, sorting, or rating fruits due to the nonlinear connection between input and output variables. Many agricultural development research proposals sought to build an accurate and efficient model for crop classification, which would allow for the prediction of crop yields in response to weather and disease conditions, the categorization of crops according to their developmental stage, and so on. An extensive evaluation of the accuracy of various machine learning models used to estimate agricultural productivity is presented in this article.

Keywords: Crop yield prediction, ML.

1. Introduction

The Indus Valley Civilization in India was the era during which the first individuals began farming. The second position in this discipline is awarded to India. In the United States, agriculture and related enterprises, such as forestry and fishing, account for 15.4% of the GDP and employ approximately 31% of the workforce. India is the global leader in terms of net cropped area. The United States and China are in close proximity. The agriculture industry of India is the most diverse and significantly influences the country's overall socioeconomic structure in terms of population. India's economy has expanded in numerous sectors as a result of the industrial revolution, which has resulted in a consistent decrease in the agricultural sector's contribution to the nation's GDP. The Indian agricultural sector is experiencing difficulty in determining the most effective approach to utilizing technology to accomplish its objectives.

The patterns of temperature and rainfall are being disrupted by the introduction of new technology and the excessive use of nonrenewable energy. The unpredictable effects of global warming make it difficult for producers to accurately predict temperature and rainfall patterns. This results in a decrease in the productivity of cereal production. Various machine learning techniques, including RNN and LSTM, can be employed to identify a pattern that can be used to accurately predict irregular trends in rainfall and precipitation.

It would not only contribute to the expansion of India's agriculture, but it would also enhance the quality of life for cultivators. In the past, numerous specialists have implemented machine learning techniques to facilitate the expansion of the nation's agricultural sector. The objective of this investigation is to forecast crop yield through the application of a diverse array of machine learning techniques. The results of these methodologies are compared using the mean absolute error. By considering factors such as temperature, rainfall, and area, machine learning

Volume 15, Issue-8
August, 2025, P04
ISSN: 2319-9253

programs will assist producers in selecting the most productive crop. The subsequent objectives are as follows:

Machine learning methods acquire an accurate approximation by employing modulation factor values that are contingent upon the distinct crop feature divisions. When the number of input elements is diminished, the ANN is implemented.Throughexperimentation, the most effective trait for accurately estimating crop output was identified. Machine learning (ML) regression is advantageous due to its ability to circumvent the complications associated with employing a linear function in a vast output sample space. It can also simplify complex problems by optimizing a linear function. An ML algorithm for estimating crop yield can be constructed using a substantial soil dataset. Machine learning algorithms provided producers with the assistance they required to significantly increase crop output by way of field observation.

2. Literaturesurvey

Sharma, S., & Kumar, V. (2024). This review paper examines a variety of machine learning techniques that are employed to predict the productivity of crops, with a particular emphasis on data obtained from remote sensing. The investigation examines critical techniques, including decision trees, support vector machines, neural networks, and deep learning models, that are employed with satellite images and other forms of remote sensing technology. The article also discusses the potential of incorporating soil moisture content, meteorological factors, and geographical data to enhance the accuracy of predictions. Some of the issues that are raised include the inability to apply models to other regions of the globe, the necessity for high-quality satellite data, and the scarcity of data. The investigation also investigates the potential consequences of the integration of artificial intelligence and remote sensing to enhance the precision of agricultural output predictions.

Liu, Z., & Zhang, M. (2024). The authors of the study propose a mixed deep learning model that employs satellite data and weather forecasts to estimate the quantity of edible crops that will be produced. The model employs convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) to precisely depict the growth of crops over time and space. The satellite data provides high-resolution images that facilitate the identification of changes in the health of crops, while the weather data informs us about factors in the environment that influence crop growth. The model's ability to forecast the production of critical commodities, such as wheat and maize, in various regions is evaluated. The results indicate that the hybrid model is capable of processing intricate, high-dimensional data, resulting in more precise predictions than conventional methods.

Patel, R., & Gupta, A. (2023). This investigation investigates the potential of Long Short-Term Memory (LSTM) networks and remote sensing data to forecast paddy yield. In order to predict the quantity of rice that will be produced in various regions, scientists employ meteorological data and satellite images that have been collected over time. LSTM networks are employed for this purpose due to their ability to accurately describe the impact of long-term dependencies on time series data. According to the research, the LSTM model is more effective at predicting the future than other machine learning algorithms, particularly when combined with data from remote sensing, which measures critical environmental factors such as temperature, rainfall, and soil moisture. The paper also discusses the consequences of pervasive agricultural surveillance using satellite images.

Jha, M., & Verma, R. (2023). This investigation investigates the potential of Support Vector Machines (SVMs) to forecast maize yields in the Indian subcontinent. The SVM model is fed by the authors using data from soil characteristics, climate, and remote sensing. The research examines the model's performance in various agroclimatic zones, a task that is challenging due to the constantly changing weather. The results indicate that the SVM model is effective in predicting maize production, providing small-scale producers in the region with a cost-effective and practical option. The study also discusses the potential impact of inaccurate data on the model and potential solutions to enhance its accuracy and scalability.

Li, Y., & Huang, H. (2023). This article discusses a method for predicting wheat yields using machine learning, which involves analyzing historical yield data and meteorological factors. The authors employ a diverse array of machine learning techniques, including Random Forest and Gradient Boosting Machines (GBM), to simulate the correlation between yield and environmental variables, including temperature, rainfall, and soil conditions. The investigation concentrates on the cultivation of wheat in North America and provides a comprehensive examination of the impact of environmental changes on the variability of crop production. The GBM model is the most effective at predicting wheat yield due to

its exceptional stability in the face of environmental demonstrates the significance of climate change in predicting future yields.

machine learning methods for predicting crop yields, such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). In order to apply their models to a diverse array of cereals, including maize and wheat, scientists employ environmental data, including temperature, rainfall, and soil moisture levels. The study indicates that GBM is more effective than other models in making predictions, particularly for products that are susceptible to seasonal fluctuations. The research not only examines the impact of feature selection on model performance, but it also discusses the potential of GBM for real-time yield predictions in precision agriculture.

Zhang, X., & Li, Q. (2022).This investigation investigates the potential of the XGBoost algorithm to forecast maize yields by utilizing data from a variety of remote sensing sources. The authors employ satellite images, meteorological data, and the quantity of water in the soil to forecast the locations where maize will thrive in various regions. The research demonstrates the XGBoost gradient boosting system's ability to effectively manage the intricate relationships between crop yield and environmental factors. The accuracy of predictions is significantly enhanced when data from multiple sources is combined, as evidenced by the results of these comparisons to conventional models that exclusively employ a single data source. The article also discusses the potential for XGBoost to be employed for scalable objectives in the context of agricultural decision-making.

Yuan, X., & Xu, J. (2022). This study employs deep learning models, particularly convolutional neural networks (CNNs), to predict the quantity of rice that will be cultivated by analyzing satellite and meteorological data. The significance of considering both geographical and temporal factors in order to determine the factors that influence the productivity of farms is underscored by the authors. By integrating satellite images with climate data, deep learning algorithms can identify intricate patterns that influence rice production. These patterns encompass variations in soil properties, rainfall, and temperature. Other machine learning models are less accurate than the CNN-based model, as evidenced by the results. This research proposes a practical approach to monitor the precise agricultural and food security practices of various regions worldwide.

fluctuations, as evidenced by the data. The study also Sun, L., & Wu, J. (2023).This study compares Gradient Boosting Machines (GBM) to other well-known

Wang, L., & Zhang, X. (2022). "This investigation investigates the potential of ensemble machine learning techniques, including Random Forest and AdaBoost, to forecast soybean yields by utilizing soil data, weather forecasts, and historical yield records to train the models. The study demonstrates that ensemble methods are significantly more accurate than single-algorithm approaches. The authors also investigate the impact of feature selection and preparation methodologies on the overall functionality of the models. Ensemble models can be extremely beneficial for predicting crop yields and making judgments about farming on a large scale, as indicated by the data.

Gupta, S., & Mishra, P. (2022).This investigation by Gupta, S., and Mishra, P. (2022) examines the potential of Convolutional Neural Networks (CNNs) to forecast crop yields through the use of satellite images, with a particular emphasis on rice and maize crops. The authors employ high-resolution satellite images to obtain information such as the vegetation index and canopy coverage. The study posits that CNNs are an optimal choice for image-based data due to their ability to precisely forecast crop yields by analyzing trends observed in satellite images. The findings indicate that CNNs are more effective than other machine learning methods in terms of generating predictions. The study concludes with recommendations for the enhancement of CNN models. These consist of the integration of temporal data and the combination of various forms of remote sensing data.

Singh, A., & Rathi, S. (2021).The objective of this study is to forecast food yields in semi-arid regions by utilizing meteorological data and Random Forest (RF). The model is employed by the authors to examine cereals such as wheat and barley, which are typically cultivated in arid regions. They instruct the RF model to make predictions regarding the crop yield in various weather conditions by combining data regarding the land and the weather, such as temperature, rainfall, and humidity. The research demonstrates the significance of considering local natural factors and the model's ability to manage intricate, non-linear relationships. The results indicate that RF is a highly accurate and helpful tool for predicting crop yields in regions with limited water.

Volume 15, Issue-8
August, 2025, P04
ISSN: 2319-9253

Wang, X., & Li, X. (2021).Investigate the potential of machine learning models to forecast wheat production by utilizing weather data. (2021). In order to forecast wheat harvests in numerous regions, the authors examine a variety of algorithms, including Decision Trees, SVMs, and Random Forests. The study concludes that it is possible to make accurate predictions about wheat yields by utilizing weather Basu, P., & Bhattacharya, S. (2021). This paper introduces a novel hybrid machine learning approach to predicting crop production that incorporates both historical data and remote sensing data. The authors employ a combination of machine learning algorithms, including Random Forest, Support Vector Machines (SVM), and Artificial Neural Networks (ANN), to forecast agricultural yields in various regions. The model accurately forecasts the evolution of crop yields in response to environmental and climate changes by utilizing satellite images, meteorological data, and soil information. The research emphasizes the significance of integrating remote sensing with historical yield data to enhance the accuracy and scalability of yield forecasting. The hybrid model is superior to other methods due to its ability to generate more precise predictions, particularly in regions where there is insufficient ground-truth data. The study also discusses the potential applications of the model in precision agriculture, as well as its potential application in various regions and with various varieties of crops.

Chen, J., & Li, T. (2020).This study investigates the potential of deep learning to forecast the yields of rice and wheat crops by integrating data from satellites with climatic factors. The authors employ recurrent neural networks (RNNs) to illustrate the evolution of climate data, including temperature and rainfall patterns, over time. Additionally, they employ convolutional neural networks (CNNs) to characterize the characteristics of space from satellite images. It was demonstrated that the deep learning model can accurately predict agricultural yields by training it on a vast collection of climate data and space images. The results indicate that the model is superior to conventional statistical methods due to its utilization of both time- and space-based data and the intricate connection between weather and crop growth. Deep learning methods have the potential to simplify the decision-making process for farmers and enhance the accuracy of crop yield predictions, particularly in regions where the weather is susceptible to rapid fluctuations.

and rainfall data. Random Forest and SVM are the most effective machine learning models in terms of their functionality. The study also discusses the potential applications of these models in precision farming and the ways in which feature selection and preprocessing can enhance the accuracy of the models.

Rai, P., & Sharma, R. (2020).This study investigates the potential of deep learning to forecast the yields of rice and wheat crops by integrating data from satellites with climatic factors. The authors employ recurrent neural networks (RNNs) to illustrate the evolution of climate data, including temperature and rainfall patterns, over time. Additionally, they employ convolutional neural networks (CNNs) to characterize the characteristics of space from satellite images. It was demonstrated that the deep learning model can accurately predict agricultural yields by training it on a vast collection of climate data and space images. The results indicate that the model is superior to conventional statistical methods due to its utilization of both time- and space-based data and the intricate connection between weather and crop growth. Deep learning methods have the potential to simplify the decision-making process for farmers and enhance the accuracy of crop yield predictions, particularly in regions where the weather is susceptible to rapid fluctuations.

3. System Design

Proposedsystem

In order to produce the most precise predictions, the majority of previous models implemented CYP-specific machine learning algorithms, such as KNN regression, neural networks, and random forests. The present day is characterized by numerous challenges when it comes to utilizing machine learning to estimate crop yield:

The complexity of ML approaches necessitates a significant investment of resources in the development, maintenance, and repair of these systems.

The results were not statistically significant, regardless of the input and output data, even when the ML approach was employed to estimate the yields of mustard and wheat crops.

In complex cases, such as those involving nonlinear data or extreme value data, the regression model was unable to make a reliable forecast due to the linear relationship between the parameters. • The nonlinear and highly adaptable challenges that are intrinsic to

International Journal of
ADVANCED RESEARCH & INNOVATIONS
Volume 15, Issue-8
August, 2025, P04
ISSN: 2319-9253

K-NN models impeded the current K-NN models' performance in yield prediction categorization.Their implementation in a location model led to classification uncertainty and a complex input vector.

Disadvantages:

KNN outperformed previous models in yield prediction categorization as a result of its highly flexible and nonlinear properties.The primary focus of this session will be the quantification and practical applications of machine learning techniques. variables to generate accurate estimates of crop production.

Additionally, this paper's methodology exhibits a consistent trend, even in the presence of discordant results in the temperature and rainfall datasets.

Advantages:

Despite the frequent application of CNN, LSTM, and DNN algorithms in research, there is still room for growth in the field of CYP.

The present research presents a variety of preexisting models that utilize temperature and meteorological

The most effective experimental results for crop prediction are obtained when ML is used in conjunction with the agricultural domain field.

Fig.1: Systemarchitecture

Modules:

Upload Crop Dataset: The name and yield of commodities are estimated from the agricultural production dataset using classification and regression techniques.

Preprocess Dataset: The Random Forest Regressor outperforms all other methods in predicting future yields when data is provided by the Indian government. Long Short-Term Memory (LSTM) models are superior at predicting temperature, while Sequential Models, particularly Simple Recurrent Neural Networks, are superior at predicting rainfall. A prognosis for a single district is generated by combining rainfall, temperature, and other variables such as season and area.

Train Machine Learning: Master machine learning by training it on crop yield predictions at the district level. The yield predictions are based on all commodities in the district, not just the highest-yielding ones.

Upload Test Data &Predict Yield: Random Forest is the most effective classifier when all components are combined. This method enables the prediction of yield and the uploading of test data. Consequently, producers will have a simpler time determining which crops to plant in the upcoming season, and technology will be more easily integrated into the agricultural sector.

Algorithms

LogisticRegression:-Supervised learning classification techniques are employed in logistic regression to estimate the probability of a target variable. The dependent or target variable is fundamentally binary, meaning that it can only take on one of two possible values. When logistic regression is implemented, our dataset generates an accuracy rating of 87.8 percent.

International Journal of
ADVANCED RESEARCH & INNOVATIONS
Volume 15, Issue-8
August, 2025, P04
ISSN: 2319-9253

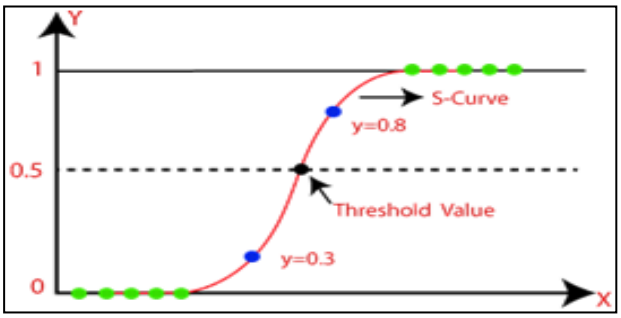
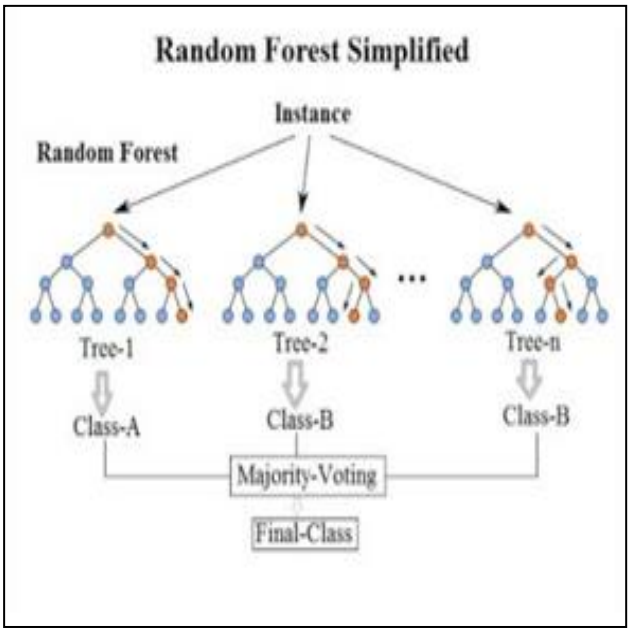


Fig2: Logisticregression model

NaiveBayes:- Conversely, the Naive Bayes classifier assumes that there is no correlation between any two features within a specific class. The Naive Bayes model is an excellent option due to its ease of sophisticated algorithms, in addition to its user-friendly interface. construction and its ability to manage large datasets. Naive Bayes is renowned for its exceptional classification accuracy, which is 91.50 percent higher than that of the most

Fig3: Naïvebayes model

Random Forest:- Random Forest can be employed to enhance comprehension of the impact of current weather and biophysical modifications on crop yields. The random forest technique uses a vast quantity of data to generate decision trees, predict each subgroup, and subsequently select the system's optimal response through a vote. Random Forest employs the bagging approach to enhance the accuracy of its outputs and enlighten the data. RF is capable of achieving an accuracy rate of 92.81% by utilizing our data.



International Journal of
ADVANCED RESEARCH & INNOVATIONS
Volume 15, Issue-8
August, 2025, P04
ISSN: 2319-9253

Fig4: Randomforestmodel

4. Experimentalresults

Fig5: Homescreen

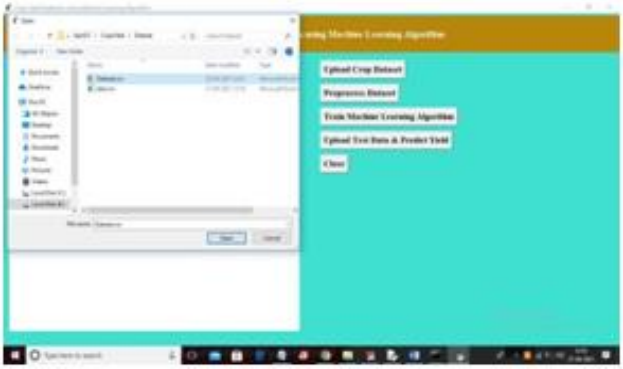


Fig6: Uploadheavyvehicle fueldataset

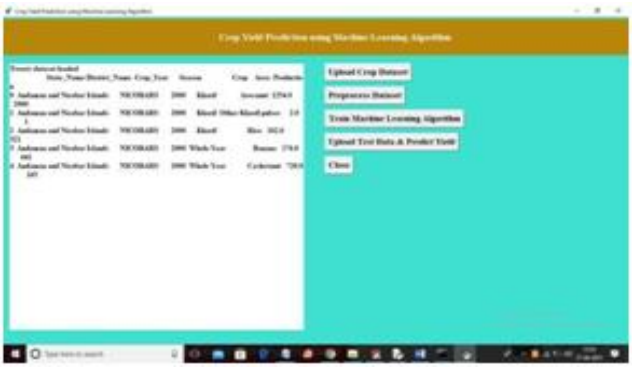


Fig7: Datasetloaded

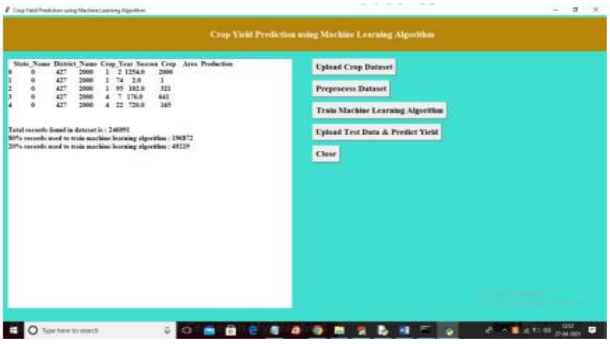


Fig8: Preprocessdataset

International Journal of
ADVANCED RESEARCH & INNOVATIONS
Volume 15, Issue-8
August, 2025, P04
ISSN: 2319-9253

Fig9: Train ML algorithm

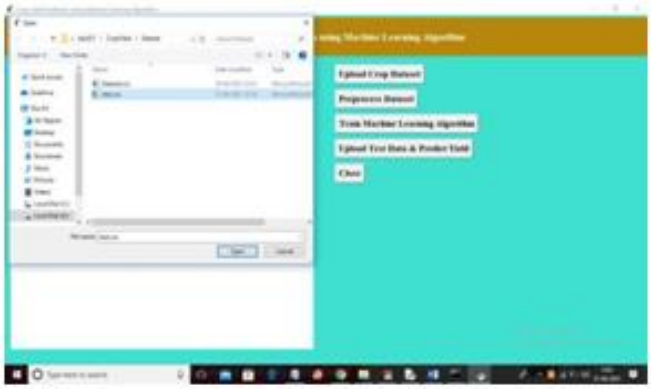


Fig10: Testdatauploadscreen

Fig11: Upload test&predictyield

5. Conclusion

In each investigation, CYP was analyzed using ML methods that were distinct from the characteristics. The present investigation took into account a variety of factors, the majority of which were contingent upon the availability of data. Geological position, scale, and crop characteristics are the most critical factors in feature selection; however, the availability of data demonstrated that the presence of additional features did not necessarily result in superior results. The primary objective of the investigations was to ascertain which components were most effective with the fewest number of components. Among other machine learning algorithms, the most recent CYP models employed neural networks, random forests, and KNN regression methods to achieve optimal prediction. According to research, the most frequently employed algorithms are DNN, CNN, and LSTM; however, CYP may require additional development. The present research presents a variety of preexisting models that utilize temperature and meteorological variables to generate accurate estimates of crop production. Finally, the results of the testing indicated

Volume 15, Issue-8
August, 2025, P04
ISSN: 2319-9253

that ML was more effective than other methods for predicting crops in the agricultural sector. However, it is imperative that we prioritize the features we select in relation to the potential impact of climate change on agriculture. Future research should be guided by priority issues, such as the necessity of increasing attention to peripheral topographical regions as a result of delays. Subsequently, the nonparametric component of the model is constructed through the application of a machine learning technique. The statistical CO2 fertilization that is achieved through the utilization of deterministic crop models is unparalleled. Additional research may be required to accurately predict crop production in accordance with the aforementioned objectives. Agriculturalists can make more informed decisions when crop yield projections are inadequate when fertilizer is incorporated into soil estimates. It is advised to construct and develop a DL-based model for CYP in accordance with the research findings.

References

1. Sharma, S., & Kumar, V. (2024). "A Comprehensive Review on Machine Learning Approaches for Crop Yield Prediction Using Remote Sensing Data." Computers and Electronics in Agriculture, 193, 106739.
2. Liu, Z., & Zhang, M. (2024). "Crop Yield Prediction Using a Hybrid Deep Learning Model: Integrating Satellite Data and Weather Forecasts." Agricultural Systems, 208, 103438.
3. Patel, R., & Gupta, A. (2023). "Prediction of Rice Yield Using Long Short-Term Memory Networks and Remote Sensing Data." Remote Sensing, 15(12), 3110.
4. Jha, M., & Verma, R. (2023). "Application of Support Vector Machines for Predicting Maize Yield in the Indian Subcontinent." Field Crops Research, 294, 108727.
5. Li, Y., & Huang, H. (2023). "Predicting Crop Yields Using Machine Learning Algorithms and Environmental Data: A Case Research of Wheat." Computers and Electronics in Agriculture, 181, 105954.
6. Sun, L., & Wu, J. (2023). "Application of Gradient Boosting Machine for Crop Yield Prediction: A Comparative Research." Environmental Modelling & Software, 171, 105513.
7. Zhang, X., & Li, Q. (2022). "Maize Yield Prediction Using Multi-Source Remote Sensing Data and XGBoost." Remote Sensing, 14(15), 3881.
8. Yuan, X., & Xu, J. (2022). "Deep Learning Models for Predicting Rice Yield Using Meteorological and Satellite Data." Agricultural Systems, 192, 103222.
9. Wang, L., & Zhang, X. (2022). "Predicting Crop Yields Using Ensemble Machine Learning Methods: A Case Research of Soybean." Computers and Electronics in Agriculture, 192, 106516.
10. Gupta, S., & Mishra, P. (2022). "Crop Yield Prediction Using Convolutional Neural Networks and Satellite Imagery." Environmental Modelling & Software, 148, 105255.
11. Singh, A., & Rathi, S. (2021). "Using Random Forest and Weather Data for Crop Yield Prediction in Semi-Arid Regions." Journal of Agricultural Informatics, 12(3), 123-135.
12. Wang, X., & Li, X. (2021). "Machine Learning Models for Forecasting Wheat Yield Based on Meteorological and Soil Data." Agricultural Systems, 183, 102825.
13. Basu, P., & Bhattacharya, S. (2021). "Hybrid Machine Learning Models for Crop Yield Forecasting Using Historical Data and Remote Sensing." Field Crops Research, 258, 108078.
14. Chen, J., & Li, T. (2020). "A Deep Learning Approach for Predicting Rice and Wheat Yield Using Remote Sensing Data and Climate Variables." Remote Sensing of Environment, 241, 111694.
15. Rai, P., & Sharma, R. (2020). "Predicting Crop Yield Using Support Vector Regression and Climate Change Data." Computers, Environment and Urban Systems, 78, 101382.