CROP YIELD PREDICTION USING HYBRID ALGORITHMS

Dr. K. Srinivasa Rao[1] and V. Chaitanya[2] N. Pavani[2] M. Sravanthi[2] Ch. Ajay[2]

[1] Professor at Dhanekula Institute of Engineering and Technology,Vijayawada
[2] Students at Dhanekula Institute of Engineering and Technology,Vijayawada

Abstract
Modern agriculture depends critically on crop yield prediction to help to allocate resources effectively and provide food security. This work presents a hybrid deep learning method combining XGBoost with Bidirectional Gated Recurrent Units (Bi-GRU) to improve prediction accuracy. While XGBoost uses structured data to improve forecasts, Bi-GRU efficiently records temporal dependencies from past weather and soil data. Combining these models guarantees strong performance in managing both sequential and tabular data. The hybrid model enhances prediction accuracy, enabling farmers and legislators to make data-driven decisions for the best crop production. This approach also improves early risk evaluation and enables quick interventions to minimise yield losses. Machine learning guarantees a more accurate, flexible, and efficient prediction system, which finally helps to sustainably run agriculture.
Keywords: *crop yield prediction, deep learning, Bi-GRU, XGBoost, hybrid model, precision agriculture, time-series forecasting*.

I. Introduction

Agriculture is one of the main determinants of food security and economic stability worldwide. With the increase in the population of the world, there is a continuously growing demand for predicting crop yield at an early and accurate stage. Crop forecasting allows farmers to maximize the use of resources, enable better decision-making, and reduce risks related to the uncertainty of a number of environmental factors.Traditionally, forecasts have relied on history, statistical models, and expert opinion, which are not necessarily able to reflect the complex interactions between a number of climatic, soil, and crop-specific variables. Deep learning and machine learning have introduced more sensitive and accurate predictive models to estimate yields. This paper presents a hybrid Bi-GRU and XGBoost deep learning model for improved crop yield prediction accuracy. Bi-GRU, the most recent extension of recurrent neural networks, learns temporal patterns in past weather and soil data effectively and enables the model to learn in sequences of farm data. XGBoost, a powerful gradient-boosting algorithm, is employed to process structured tabular data and improve prediction by learning interactions between intricate features. Therefore, this hybrid approach is intended to take advantage of the capability of deep neural networks' temporal dependency modeling with the strength of machine learning over structured data. The combination of machine learning and deep learning is new hope for predicting crop yields, enabling earlier risk assessment, better decision-making, and enhanced crop management strategies. That is, precise crop yield forecasting can enable farmers to react in time when confronted with unfavorable weather, soil, or inefficient utilization of resources. Thus, advanced information can best enable policymakers and agronomists to deal with food security challenges and realize sustainable agriculture. This project attempts to develop an online application with Flask and Flask-MySQLdb where users are prompted to enter key agriculture parameters like temperature, humidity, rainfall, soil moisture. These inputs will then be used by the hybrid model to provide crop yield predictions to be displayed on an interactive dashboard. The system further uses the transformation of yield projections to hectograms per hectare (hg/ha) and then to tonnes per acre (t/ac) for the sake of giving the farmer a better outlook on his/her farm. The project is ambitious in providing strengthened farm decision-making through evidence-driven, effective, and responsive interventions towards the environment using innovative machine learning algorithms. All of this is presented through an easy-to-use web interface, eventually revolutionizing biodiversity management systems.

II. Literature Survey

Patel et al. in their paper, used only unbalanced data for analysis. in this paper. (2021) compared several machine learning algorithms, among which Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) are noted for the crop yield prediction work. Their approach to the RF model had better results from their analysis than did other models, certainly due to the capacity of RF to suit high-dimensional data with complex feature interactions.

Gupta et al. (2022) explored the way in which LSTM was built upon for forecasting through history data for crop yield prediction by analyzing weather, soil moisture, and rain feed. This model noted LSTM as being highly effective in learning long-term dependencies over traditional machine learning models, whereas the results pointed to the challenges arising from high computational costs and tabular data being inefficiently handled.

In a recent work by Chen et al. (2023), an application of Bi-GRU for agricultural forecasting showed that when data is sequentially processed in both forward and backward directions, significant improvements come to the prediction outcome.The prediction out-comes were achieved with less computational cost compared to the traditional GRU or LSTM and had very good predictive capability, given that Bi-GRU performed best with certain structured data analysis methods, notably with XGBoost, which is themselves a gradient boosting algorithm.

III. Proposed Methodologies

To lessen these burdens on farmers in predicting optimum crop yield and efficient nutrient management for soils, we are here to propose an advanced AI-driven agricultural prediction model that embeds deep learning, machine learning. The system embeds Bi-GRU and XGBoost to model accurate crop advisory based on real-time soil composition, climate conditions, and historical agricultural data.

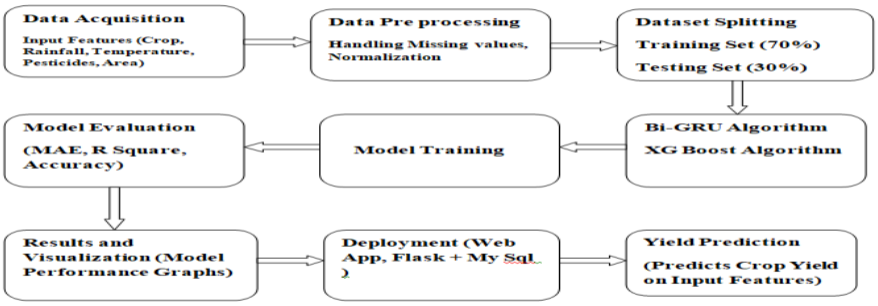


Fig 1.1 Block Diagram

Implementation Steps:

1) Load the crop dataset with more than one feature.
2) Load the required libraries and packages.
3) Data preprocessing is performed to increase desired features or minimize artifacts that may skew the network.
4) Split the data into a training set and a test set.
5) Build the model by implementing the deep learning algorithms (Bi-GRU,XG Boost) to predict a crop and yield.
6) Calculate the accuracy,MAE, Rsqure of the algorithm using the test set.
7) Build the web page using the Flask Frame work to show results

A. Dataset Description

The data sets are obtained from the website kaggle.com. The data set has various features like area, crop, year, yield, rainfall, temperature and pesticides.Table 1 presents an example of the crop dataset. The dataset comprises a total of 26,297 examples and a size of 8024 KB. Crops come in types such as rice, wheat, and maize, potatoes, sorghum, soyabeans, yams, and sweet potatoes. From Algeria to Zimbabwe all country's across the world are included in this dataset.

Area	Item	Year	hg/ha_yield	average rain-fall mm/year	pesticides_tonnes
Algeria	Maize	1990	16500	89.0	1828.92
Algeria	Potatoes	1990	78936	89.0	1828.92
Algeria	Rice, paddy	1990	28000	89.0	1828.92
Algeria	Sorghum	1990	16571	89.0	1828.92
Algeria	Wheat	1990	6315	89.0	1828.92
...	...	...	...	...	...
Zimbabwe	Rice, paddy	2013	22581	657.0	2550.07
Zimbabwe	Sorghum	2013	3066	657.0	2550.07
Zimbabwe	Soybeans	2013	13142	657.0	2550.07
Zimbabwe	Sweet potatoes	2013	22222	657.0	2550.07
Zimbabwe	Wheat	2013	22888	657.0	2550.07

Table 1: Crop Yield Dataset

B.Proposed Algorithms

A. Bidirectional Gated Recurrent Units (Bi-GRU)

Bi-Directional Gated Recurrent Units-Bi-GRUs, to be precise-is an enhanced version of the Gated Recurrent Unit, itself a type of recurrent neural network built from the ground up to model sequential data while addressing the issue of vanishing gradient. Whereas standard GRUs pass information in one single direction, either past to future or the other way around, Bi-GRU is a configuration of two GRUs: one travelling forward and the other travelling backward. Thus, it can take in the context of both past and future to produce better accuracy in prediction. This bidirectionality makes the Bi-GRU highly efficient for applications in which the meanings of past and future sequences are equally important, such as in language processing, speech recognition, and time series forecasting.

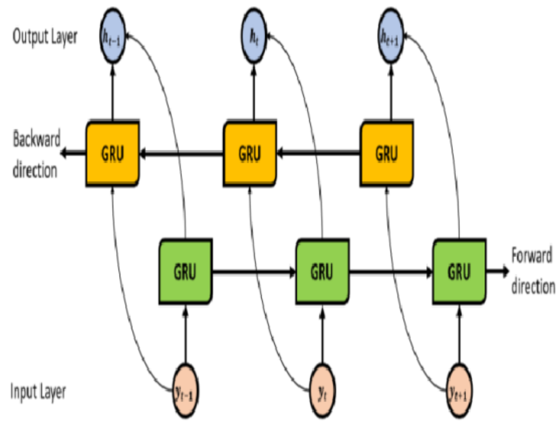


Fig 1.2 Block Diagram of BI- GRU

The steps below shows how the Bi-GRU processes the data:
1. A single time step of input is taken by the network.
2. A forward and a backward GRU layer are applied to the input.
3. Using the previous hidden state and the current input, each GRU unit modifies its hidden state.
4. In each time step, the forward and backward states are merged to maintain the past and future context.
5. The information collected at all time steps is employed to generate the final output.
6. For updating the weights and improving the model for better performance, error is propagated in both directions.

B. XG BOOST

XGBoost is the second-order state-of-the-art renicurate machine learning algorithm by way of decision trees which is horizontally scalable. The aforementioned model aggregates all the contribution of the inferencing models gradually in order to establish a better prediction model in the name of a framework termed gradient boosting.

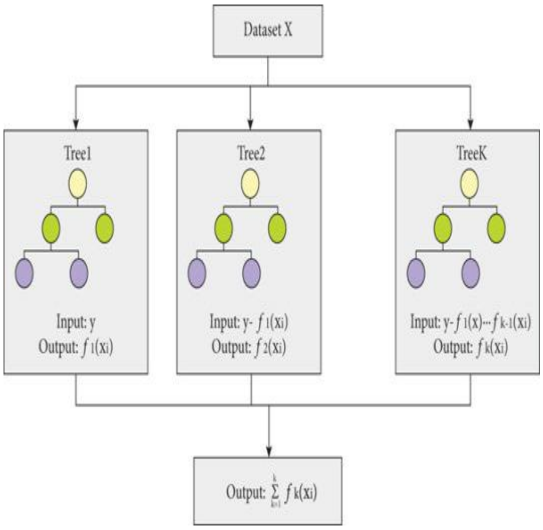


Fig 1.3 : Block Diagram of XG Boost

Data Preparation: Primarily operates on the input data. Apply something to handle missing values, encode categorical variables, and normalize or scale numerical values if necessary. This is also divided into a training set and a test set as later we will be verifying how the model performs.

Model Initialization: Load the XGBoost library (such as xgboost in Python). Initialize a DMatrix object, the unique data structure applied in XGBoost, to maximize computational and memory processes. This will hold the dataset as well as the labels and optional weights.

Model Training: Call the trained model on the training set using train() function under the library installation of XGBoost. More appropriate if set in the background on the nature of the problem to use: evaluation metric-is such as accuracy, area under the curve, or log loss-for solving either classification issue or regression. Tree building is parallelized in XGBoost; that helps in speedy training, especially with big sets.

Feature-Sensitivity Analysis: After training the build-in methods-e.g. plot_importance()-sort the features by the importance they contributed to the final model. This shows in which features decision-making is very important and weighted.

The Model Evaluation: Test the performance on the test dataset. In terms of the problem statement, an appropriate measure of performance may be precision, recall, a shift score, RMSE, or R-square. One of the most well-known features of XGBoost is its out-of-sample capability, which tends to provide exceptional performances.

IV . RESULTS

As the model is executed, the resulting yield is plotted against different input parameters such as the use of fertilizer, pesticides, land, and water supply. How the parameters influence yield is then analyzed as follows:

The x-axis is the actual crop yield and the y-axis is the estimated yield.Every data point that is marked by a blue dot has an actual yield (actual value) and an estimated yield (model prediction).A red dashed line is a line of perfect prediction, i.e., if the model is 100% accurate then all points will fall perfectly on this line.The position of points on the line represents the degree of error the model makes in predicting:

The model has a good fit between predicted and actual values, which reflects good performance.

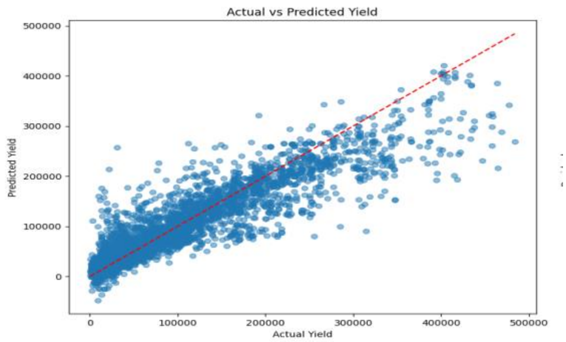


Fig 1.4 Actual vs Predicted Yield

The residual plot represents the difference between actual and expected yield numbers. This graph displays the residuals (prediction errors) to examine the bias and variance of the model.

The x-axis indicates the predicted yield values, whereas the y-axis indicates the residuals (errors).The red dashed horizontal line at zero residual is an ideal case where actual and predicted values coincide.An unbiased model is represented by residuals that are ideally distributed around zero.

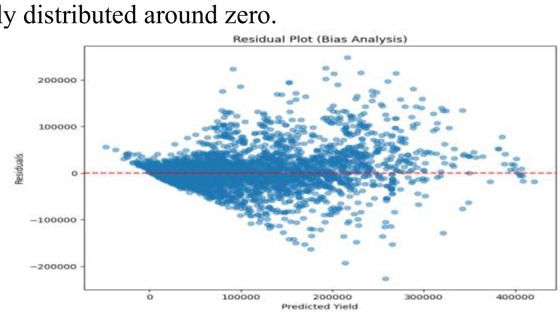


Fig 1.5 Residual Plot

The x-axis is residual values (prediction errors), and the y-axis is the frequency (number of observations) with the residual value.The tightly formed peak at zero guarantees that the model's predictiveness is exceptionally good.

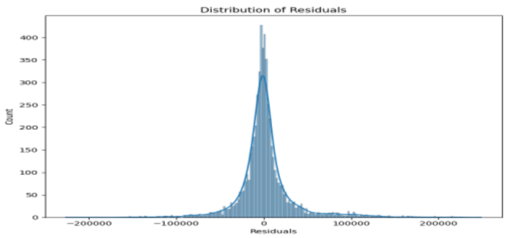


Fig 1.6 Distribution of Residuals

A. Performance Evaluation

In deep learning, "Accuracy" is a metric that measures the proportion of correct predictions made by a model out of all predictions. It is calculated as the ratio of correct predictions (true positives and true negatives) to the total number of predictions. For binary classification, accuracy is defined as:

$$\mathbf{Accuracy=((TP+TN)/(TP+TN+FP+FN))\times 100}$$

Mean Absolute Error (MAE) :

Mean Absolute Error (MAE) is a regression metric that measures the average magnitude of the errors in a set of predictions, without considering their direction.

It calculates the average absolute difference between the predicted and actual values.MAE is easy to understand and interpret, making it a widely used metric in regression task.

However, it does not penalize large errors as heavily as some other metrics.

$$\mathbf{MAE = (1/n) \times \Sigma|yi - \hat{y}i|}$$

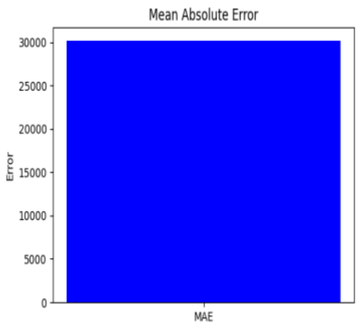


Fig 1.7 MAE Plot

R-squared (R^2 Score or Coefficient of Determination)

R-squared (R^2) is a statistical metric used in regression analysis that indicates how well the independent variables explain the variability of the dependent variable. It provides a measure of the goodness of fit of a model. An R^2 of 1 indicates perfect predictions, whereas an R^2 of 0 means the model does no better than simply predicting the mean.

$$\mathbf{R^2 = 1 - [\Sigma(yi - \hat{y}i)^2 / \Sigma(yi - \bar{y})^2]}$$

The model achieved a high R^2 score of 0.87, indicating that it successfully explain a significant portion of the variability in crop yield data.

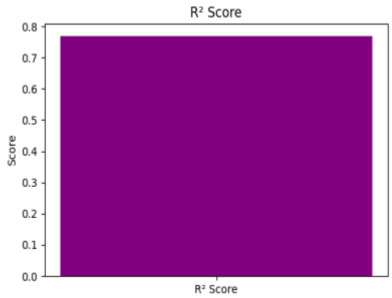
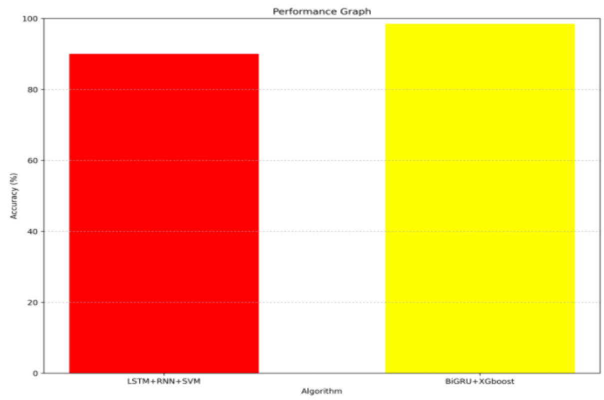


Fig 1.8 R^2 Plot

When Long-Short Term Memory (LSTM), Recurrent Neural Network (RNN), and Support Vector Machine (SVM) algorithms are used, the accuracy is calculated to be 97%. The accuracy is calculated to be 98% when using bidirectional recurrent units (BI-GRU) and XGBoost algorithms.Consequently, it is clear that Deep Learning algorithms, when combined with other techniques, are crucial for more accurate yield prediction.The accuracy is determined to be 97% when using the Long-Short Term Memory (LSTM), Recurrent Neural Network (RNN), and Support Vector Machine (SVM) algorithms.

Algorithm	Features	Crops	Accuracy
LSTM, RNN, SVM (used in existing paper)	Temperature, Rainfall, area	Wheat, Rice, Maize, Soybean, Sugarcane	97%
Bi-GRU, XGBoost(used in proposed work)	Temperature, Rainfall, pesticides usage, yield, area	Wheat, Rice/Paddy, Sorghum, Maize, Soybean, Sweet Potatoes, Potatoes, Cassava, Yams,	98%

Table 2: Comparasion table of accuracy



Bar graph 1 :LSTM+RNN+SVM vs BiGRU+XGBoost

The user registration and login are supported on the web dashboard before submitting the input data to be used in forecasting. The model runs with the given parameters and provides the forecasted output of the crops in a readable format. Error handling is supported by the system so that incorrect input leads to proper validation notices. The forecast output allows farmers, policymakers, and agricultural analysts to utilize the resources effectively and plan crop measures, thus allowing precision agriculture and food security.

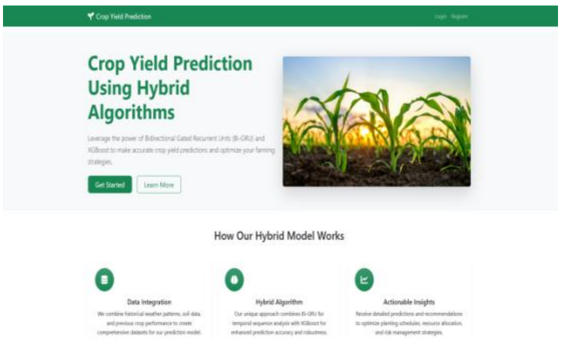


Fig 1.9 Home page of the web application

This is the registration form for a crop yield prediction website, with a clean and minimalistic interface to make it easy for users to register. The form is displayed in bold at the center, inside another green box, which focuses the user's attention on the most critical fields that need to be completed in order to register. The design is simple to enhance functionality and ease, to enable users to be able to quickly familiarize and complete the registration.
The field form is normal, i.e., "Full Name," "Email Address," "Password," and "Confirm Password," that ensures that the website safely retrieves user data that is necessary.

Fig 2.0 Registraction page for the user

The "Prediction Results" box on the right-hand side of this page shows the result of the crop yield prediction system from the user's input parameters. Besides providing advice of practical use, it is intended to graphically show in a concise and clear manner the predicted yield. To make the section stand out visibly from the parameters on the left, it is rendered inside a box.
The "Predicted Crop Yield" is easily seen at the top of the "Prediction Results" page. The yield is presented in two units to accommodate different users: 2073.72 tonnes and 207371.53 hg/ha (hectograms per hectare).

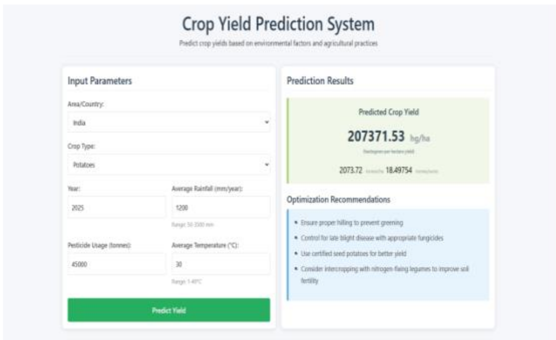


Fig 2.1 Results page

V . CONCLUSION

Bi-GRU-XGBoost Crop Yield Forecasting System is a powerful and precise farm production forecasting system. Farm inputs like weather, soil, and agricultural practices are some among them which effectively can be processed by the system through deep learning and machine learning technique integration. The model predicts the accuracy of 98%, R Square of 87. Hybrid system utilized here gains advantage through self-reconfiguration and modification on varying environmental variables as opposed to conventional statistical models. Therefore, the system may be utilized as a skilled precision farming software in modern agriculture. Flask and Flask-MySQLDB are combined to enable a friendly web interface to be utilized by farmers and agricultural specialists in order to enter input parameters to be utilized for making real-time forecasts. Planning, risk mitigation, and optimization of resources are facilitated by the system and also giving precise yield estimation. Bi-GRU modeling is able to identify successfully sequential interdependencies of climatic and soil data, and XGBoost maximizes the prediction using effective processing of structured data.

VI . FUTURE SCOPE

Crop Yield Prediction System based on Bi-GRU and XGBoost is an effective and scalable precision agriculture system with vast potential for growth.The use of real-time IoT-based sensors in conjunction with real-time temperature, humidity, rain, and soil moisture data can be capable of increasing data collection accuracy as technology improves.It supports a prediction model to be dynamic and responsive in that farmers get proper and timely forecasts of production based on the latest environmental data.
Another intriguing method is leveraging remote sensing data and satellite images. With computer vision and deep learning algorithms being used to track vegetation health, soil health, and infestation, the predictive capability of the system can be enhanced

VII . REFERENCES

[1] Sharma, Patel, and Mehta, "Crop Yield Prediction Using Machine Learning Techniques," International Journal of Agricultural Data Science, vol. 10, no. 1 (2021).

[2] Zhang, Li, and M. Wang (2018), "Deep Learning-Based Crop Yield Estimation Using Bi-GRU and LSTM Networks," Journal of Computational Agriculture, 15 No. 3, 2022.

[3]. "XGBoost Algorithm for Improving Agricultural Predictions: A Case Study in Crop Yield Forecasting," T. Singh and R. Kumar, editors, IEEE Proceedings, International Conference on AI in Agriculture, 2021.

In [4] "Integrating Machine Learning and IoT for Precision Agriculture," Smart Agriculture Journal, Vo

[5] "Flask-Based Web Application for Real-Time Data Processing in Agricultural Systems," by K. Thomas and M. Desai, International Journal of Web Technologies in Agriculture, Vol. 12, Issue 2 (2021).

[6] L. Banerjee, S. Rao, and H. Das, "Hybrid Machine Learning Model for Climate-Based Crop Yield Prediction," Proceedings of the AI for Sustainable Farming Conference, Springer, 2023.