Measurement: Sensors 35 (2024) 101277

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

Ensemble regression based Extra Tree Regressor for hybrid crop yield prediction system

T. Sudhamathi [a,*], K. Perumal [b]

[a] Department of Computer Applications, School of Information Technology, Madurai Kamaraj University, Madurai, India
[b] Department of Computer Applications, School of Information Technology, Madurai Kamaraj University, Madurai, India

ARTICLE INFO

Keywords:
Crop yield prediction
LESSO regression
Kernel
Principal component analysis
Ensemble regression
Extra tree regressor

ABSTRACT

Objective: The worldwide economies are built on agriculture, and plans for food security, resource allocation, and agricultural practices are all heavily influenced by accurate crop production predictions. Predictive models are becoming indispensable tools for predicting crop prospects due to the development of technology based on data.
Limitation: A significant disadvantage of the ER-ETR for Hybrid Crop Yield Prediction System can involve overfitting, particularly in cases when the dataset is small or the model complexity is not well managed. Inaccurate forecasts based on unreported data and decreased generalization can result from approach.
Method: Initially, the dataset is collected from the GitHub and preprocessed using the Standardscaler method. 70 % of the preprocessed data is used as the training set, and the remaining 30 % is used as the testing set. Kernel Principal Component Analysis (KPCA) is employed to extract the feature. The Least Absolute Shrinkage and Selection Operator (LESSO) Regression is used to feature selection.A reliable method for predicting hybrid crop productivity is provided by the suggested ensemble regression that makes use of feature ensemble regression using Extra Tree Regressor (ER-ETR).
Result: A simple internet-based programme for immediate forecasting is created using the Python web framework, and the model that has been trained may be used to predict the resulting profitability. Mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE) and R^2 were the testing metrics utilized to assess the classification model. With a 95 % accuracy rate, the suggested model is superior to existing models in terms of accuracy in crop production forecasting while still preserving the data's original distribution.Because of the intuitive online interface, stakeholders can forecast immediately and make well-informed decisions on the best use of resources from agriculture.
Conclusion: The study creates a hybrid crop yield prediction system using the ER-ETR approach. Agricultural forecasting benefits greatly from its capacity to integrate several models and take advantage of each one's advantages, which improves prediction accuracy and dependability.

1. Introduction

One of the significant professions practiced in the country is cultivation. By engaging in numerous agricultural operations, the country's finances significantly improve. Consequently, it is referred to be the most comprehensive revenue earning approach. For agricultural purposes, 60.45 % of the land in India is exploited. Doing so helps 1.2 billion people get what they need. Nowadays, there is a significant effort being made to modernize the agricultural industry. As a result, the farmers are positioning themselves for success and making more money with less outlay. Data science is used to evaluate informational indexes, guiding the use of specific tools and frameworks as well as the ability to conclude the data they include. Historically, a land owner's familiarity with a specific plot of property and crop served as the foundation for production estimation. As conditions gradually alter, producers focus on amassing an ever-increasing quantity of commodities. A lot of agricultural producers are interested in learning more about the higher revenues in light of the current situation (Agarwal and Tarar [1]). Forecasting crop yields is critical to ensuring that there is an adequate supply of food and boosting the nation's overall level of food security. It is crucial to do this assignment at the regional and national levels with extreme precision in order to make decisions quickly. For instance, when

* Corresponding author.
E-mail address: sudhamathit@hotmail.com (T. Sudhamathi).

https://doi.org/10.1016/j.measen.2024.101277
Received 22 March 2023; Received in revised form 29 March 2024; Accepted 7 July 2024
Available online 9 July 2024

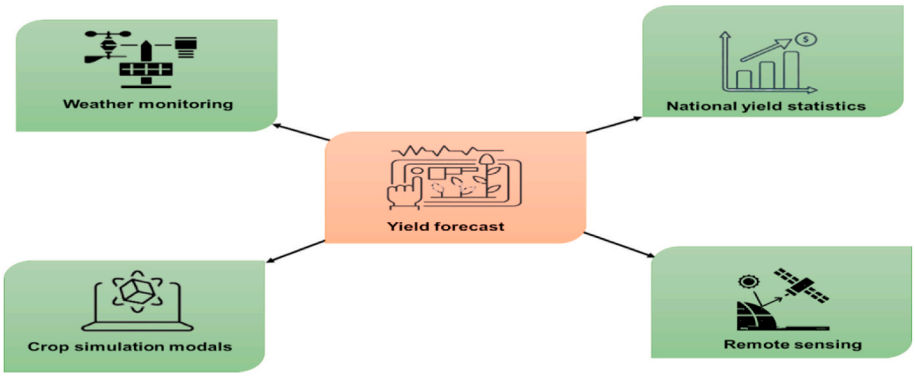

Fig. 1. Crop yield estimation method.

the outcomes of the agricultural production forecast are correct, officials can make judgments about exports and imports. This kind of forecast helps farmers make sound financial choices. Additionally, seed producers can assess how new seeds perform in various settings Kim et al. [2].

Numerous programs have been launched in recent years to reduce world hunger and feed the planet's expanding population. Nearly 800 million people still lack enough food to consume despite a significant rise in agricultural productivity over the last 50 years. The yield estimate techniques employed in this technological age for various crops are shown in Fig. 1. The most well-known of these are remote sensing, Crop Simulation Models (CSM). These methods are also used to determine how the climate affects the agricultural output of various varieties (van et al. (2020)). For many players engaged in the production and trade stage of agriculture, predicting the crop's potential yield is an important milestone. To help farmers control their resources and resource consumption, giving them an estimate of their output is essential. By doing this, farmers are better equipped to make managerial and economic decisions, and early detection of problems that impact productivity can help with timely corrective action for the entire crop.Estimating the yield of crops can enable better-informed decisions to be made while organizing and carrying out tasks. This makes predicting agricultural production a complex task that has to be resolved. Several factors have an impact on crop output levels.Including seed variety, soil and weather conditions, fertilizer usage, and soil conditions, which also affect plant phenotype Xu et al. [3]. This research suggests using Ensemble Regression-based Extra Tree Regressor (ER-ETR), a hybrid deep learning and machine learning technique, to forecast the crops. Our study intends to give guidance to both academics and industry professionals on how to employ ML and DL approaches for agricultural production prediction, as well as potential solutions. PCA is used for feature extraction.

1.1. Contribution of the study

- To forecast crops, we created hybrid machine learning and deep learning technique termed Ensemble Regression-based Extra Tree Regressor (ER-ETR).
- For feature selection, we apply Kernel PCA. We evaluated the ways in which the performance of different feature techniques was impacted by the number of crops prediction approach.
- We showed that among the technique, the ER-ETRtechnique performs the best, while the KPCA model comes in second.
- The Standardscaler technique is first used to get the dataset from GitHub and preprocess it. The training set is comprised of 70 % of the preprocessed data, while the testing set is made up of the remaining 30 %.

The next part of this report is structured as follows: An overview of the manuscripts is given in Part II. In Part III, the procedure is explained. Part IV covers the experimental results and discussion, while Part V offers the conclusion and recommendations for more study.

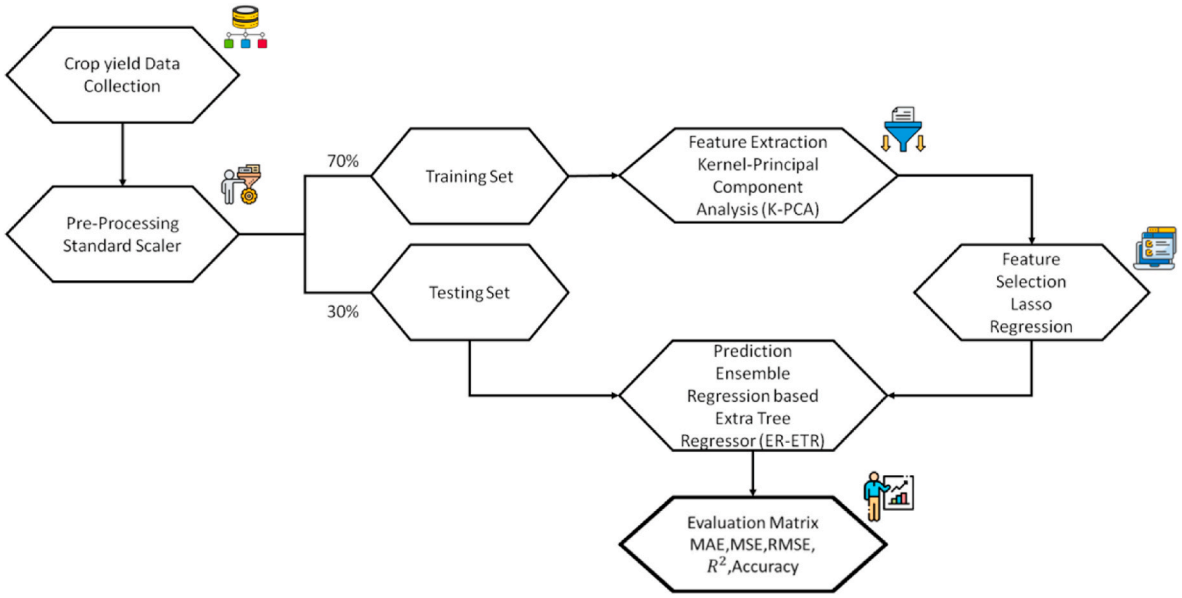

Fig. 2. Architecture of Proposed PCA with ER-ETR model.

Author	Data	Results	Benefits	Drawbacks
Batool et al. [4]	Data on weather, soil, crops, and agro-management were gathered from tea fields at Pakistan's National Tea and High-Value Crop Research Institute (NTHRI) between 2016 and 2019 to train regression algorithms and calibrate the AquaCrop simulations model.	The ML can accurately estimate agricultural production, surpassing conventional techniques, as demonstrated by the effective prediction of tea output using a deep neural network approach.	The ML and hybrid techniques enhance agricultural forecasting accuracy, aiding producers in decision-making and food security, while DNN techniques and improved programming methodologies facilitate successful forecasts in intricate agricultural systems.	ML-based agricultural prediction systems face challenges like biases, data limitations, and sampling mistakes, necessitating detailed analysis of data mistakes and uncertainties for improved effectiveness.
Cao et al. [5]	The study to predict winter wheat yield at the grid level from 2005 to 2014. They also used three major datasets: satellite data from MOD13C1, qualitative climate information from CRU, and S2S atmospheric prediction data from IAP CAS.	The findings also demonstrated the effectiveness of MHCF v1.0, which had the greatest R2 in winter wheat yield forecasting.	For yield forecasting, agricultural practice guidance, policy guidance, and agricultural insurance, the ML-S2S hybrid model proved to be a valuable device.	The hybrid ML-dynamical weather forecasting system may face limitations due to insufficient investigation of potential data bases or ongoing model modifications for precise forecasting under various scenarios.
Khaki and Wang [6]	In the 2018 Syngenta Crops Challenge, participants were asked to forecast the yield performance in 2017 based on a number of sizable datasets that included information on the genotype and yield performance of 2267 maize hybrids planted in 2247 sites between 2008 and 2016.	The DNN model outperformed other models in forecast accuracy, with an RMSE the average return and 50 % of standard deviation, and improved forecast accuracy when ideal weather conditions were present.	The DNN technique demonstrated the potential of advanced algorithms in agricultural forecasting by enhancing yield prediction accuracy and model efficiency by reducing input dimensionality.	Depending on huge databases and precise meteorological data. A lack of consideration was paid to potential biases or data uncertainties that could have an impact on how well the model performs in practical situations.
Colombo-Mendoza et al. [7]	A climate dataset comprising the following observations made in the municipalities of the northeastern section of the Mexican State of Puebla between 2003 and 2019 was gathered.	The smart farming system incorporates data-mining and IoT technology to improve crop output forecasts, aiding peasant farmers in decision-making through historical data validation.	IoT and data mining enhance agricultural decision-making, benefiting small-scale farmers with affordable solutions, enhancing scalability and accessibility through data analytics and cloud services.	The system's dynamic adaptability to changing environmental conditions may be hindered by its reliance on past data and traditional machine learning methods, and potential scalability and data protection issues.
Feng et al. [8]	The Scientific Information for Land Owners patched point dataset (SILO-PPD) included the 29 locations' recorded daily climate data (2008–2017), which included details on solar radiation, precipitation, and the lowest and highest air temperatures.	The results of regression analysis will be used to develop a hybrid yield forecasting approach.	The method identifies drought as the main cause of production losses, providing accurate crop output projections for farmers and decision-makers, potentially expanding to yield projections with more farming data.	The study did not address potential limitations or difficulties in the design and application of the hybrid yield prediction method, such as data quality or scalability issues.
Shah Hosseini et al. [9]	A data collection comprising observed data on yields, management, and the environment was created. It consists of 10,016 samples of the average annual maize yields for 293 counties.	The results demonstrate that a certain creative characteristic may provide a detailed explanation of trends and is essential for predicting maize yields.	The study provides insights for enhancing prediction reliability by assisting in the identification of the most efficient model and feature combinations for precise maize yield forecasts.	Potential drawbacks include the necessity for additional validation in actual agricultural settings to verify the efficacy and dependability of the suggested CSM + ML approach for maize yield prediction
Cooper et al. [10]	The possible grid-level uses of Digital gap analysis (DGA) examples. Any of the 2282 grids that represent the U.S. corn-belt Target Population of Environments (TPE) may have its simulated findings analysed using the same gap analysis method as the empirical study mentioned above.	The GEM yield forecasts from CGM are tested in three scenarios: yield potential, drought studies, and hybrid maize genetic gain, highlighting their relevance for crop development initiatives.	The CGM framework enables swift development and testing of crop improvement plans utilizing GEM interactions, thereby enhancing productivity and closing yield gaps.	The study does not investigate potential challenges in implementing the CGM paradigm in real agricultural settings, such as scalability, model complexity, or data availability.
(Anbananthen et al. [11])	A novel model called stacked generalisation discovers the optimal way to integrate forecasts from multiple models that have been tested on the dataset. The aerial-intel datacollected from the github data analysis repository serve to illustrate how the suggested approach may be employed.	Cross-validation results demonstrate how well individual and hybrid machine learning algorithms perform in comparison. Accuracy is attained using stacking generalisation, gradient boosted tree, and random forest.	The stacking generalization ML method effectively predicts crop production, responding promptly and accurately to farmers' needs.	The real-world agricultural settings may face limitations due to inadequate examination of potential obstacles like scaling issues, computing resource constraints, and data availability.
Keerthana et al. [12]	Data collection involves removing irrelevant columns from various datasets to obtain food and agriculture data from github and Kaggle, resulting in 7 columns and 28242 occurrences.	The finding in Decision Tree and AdaBoostregressor ensemble effectively forecast crop production, providing advice on crop cultivation based on local weather conditions.	The study demonstrated the effectiveness of ensemble machine learning methods in improving agricultural yield forecast accuracy, simplifying decision-making, and demonstrating the potential for better predictions through multiple ML methods.	The insufficient thoroughness of data analysis in ML models can significantly affect the precision and dependability of real-world crop production projections.

2. Literature Survey

2.1. Problem statement

They must develop a system of recommendations that can propose an enhanced crop that would provide the best yield. Utilizing machine learning methods to anticipate agricultural production is the primary goal of the problem description. The project's goal is to aid users in making the optimum crop choice to optimize yield and, subsequently, profit. Structured data analysis is used to produce predictions to get over the restrictions of the proposed method. The method we recommend is to develop a system that takes into consideration the elements that have the most impact on how well a crop develops and to broaden the range of crops that can be planted throughout the season. For many participants in the production and trade stages of agriculture, predicting the crop's future yield is a significant accomplishment. To help farmers plan their budgets and manage the usage of resources, it is crucial to equip them with yield prediction systems and future problem-solving strategies.

3. Proposed Methodology

Deep learning is a machine learning technique that uses artificial neural networks to extract intricate patterns from large data sets, enabling advanced tasks like voice and image recognition.Deep learning is the process of learning complex patterns by training neural networks with several layers. Deep learning algorithms have the ability to evaluate large agricultural datasets and include many aspects, such as crop genetics, weather, and soil conditions, to estimate crop yields more accurately and efficiently.In the suggested structure, ML and DL techniques are used to estimate the optimal yield of crops. The suggested model runs experiments using an inventory of vegetables. The crop is chosen based on the current environment, the ground, and its components, while taking meteorological and soil characteristics into consideration. DL is employed to do a wide range of useful mathematical such as selecting the optimal crop when there are several options. Accurate forecasting of crops is possible using this strategy. Accurate agricultural production forecasting can be greatly improved by utilizing DL in ER-ETR. The goal is to offer more accurate forecasts that are essential for agriculture planning and decision-making by combining the adaptability of DL frameworks with the stability of tree-based models.The ER-ETR paradigm is used to construct a hybrid ML and DL technique, as shown in Fig. 2.

3.1. Dataset

GitHub is software tool that, when used with the program known as Git, enables version management of your various projects. It gives the users the ability to disseminate the code, collaborate on projects, assist in the deployment of projects using CI/CD (continuous integration/continuous deployment), revert to prior versions of the project, and a variety of other capabilities. In this part, we will show our replication package and discuss the approach that we used to obtain our data. GitHub provides yield df.csv support for projects that have access to the functionality and told us that they were unable to release the information when we first started this study endeavor. We had no choice but to depend on a web scraper that we had developed specifically for the purpose of determining which projects were already in use to gather information and postings for analysis. There are a total of 115 columns and 5 rows in the data to be tested.Table 1 displays this exemplar gathering of information.Using APIs or databases that offer historical and current weather forecasts for the pertinent geographic locations is one way to include meteorological data with the prediction model. When combined with additional work data from GitHub, this data might be incorporated into the current dataset to analyze the influence of meteorological factors on crop yield forecast.

3.2. PreprocessingusingStandardscaler

Standardization, often referred to as standards scaling, is a step in the preprocessing of the data. Within a specific range, it is used to normalize the data. Measuring features facilitates faster algorithmic computations. Variables with various criteria make up the dataset utilized in this experiment. Numerous bars are available for measuring the dataset. Regularly used scalars include robust scalar, min-max scalar, normalizer, and standard scalar. Since StandardScaler is used to transform datasets, the resultant distribution's average is zero, and its standard deviation is one. To obtain the converted value, take the original number, remove the average from it, and then divide the result by the standard deviation. To adjust, utilize the formula below.

$$Y = \frac{W - \mu}{\sigma} \tag{1}$$

Where σ is the standard deviation, μ is the mean, and the initial value is W. Preprocessing techniques for data, such as scaling, conversion, and feature reduction, create a standard dataset and lower prediction errors. The effectiveness of models is impacted by choosing attributes, scaling, data cleansing, and trimming. The crop yield prediction depends heavily on standardization in data preparation, which may be achieved using tools like StandardScaler. By normalizing data within a predetermined range, its procedures and guarantees uniform measurements across variables. For crop yield prediction models, this preprocessing stage provides an appropriate number of crops from dataset.

3.3. Feature extraction usingKernel Principal Component Analysis

The process of creating numerical features from raw, unedited data without compromising the integrity of the original data set is known as feature extraction. In the world of data science, a common technique for feature extraction isPrincipal Component Analysis (PCA) supported by kernel OS is called it us kernel PCA. PCA re-projects the data into a subspace with the same number of dimensions or less by using the biggest eigenvectors of a covariance matrix. Using the idea of dimensionality reduction, it divides multi-indications into a number of specific indicators. Therefore, II hope that you can incorporate a few more elements and gather more data if the research going to undertake a quantitative analysis of variables. The more variables that are studied, the more computationally intensive and challenging the problem analysis becomes. For this objective, principal component analysis is the most effective technique. Since the purpose of the main component is to minimize dimensionality and, as a result, the number of variables, a limited number of principal components are frequently used in practical research as long as they can hold more than 80 % of the information of the original variables. This is so because the primary component's major objective is to reduce dimensionality.

To minimize the effects of dimensions, standardize raw data:

Assuming there are n objects, define x_{i1}, x_{i2}, …, and x_{ip} as the i-th object's corresponding p indexes. Use the matrix (2) below to represent all observations of the p indices of n objects.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{1p} \\ x_{21} & x_{22} & x_{2p} \\ \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & x_{np} \end{bmatrix} \tag{2}$$

Where n = No. Of objects and.

p = No. Of indicators/variables.

Keep applying the standardization approach to the data of the p indices of the n items in accordance with the following equation (3).

$$x'_{ik} = \frac{x_{ik} - \overline{x}_k}{S_k};\ \ i = 1, 2\ldots, n\,; k = 1, 2, \ldots, p \tag{3}$$

Table 1
Example information set.

Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp	Area_ Albania	Area_ Algeria	Area_ Angola
0	1990	36613	1485.0	121.0	16.37	1	0
1	1990	66667	1485.0	121.0	16.37	1	0
2	1990	23333	1485.0	121.0	16.37	1	0
3	1990	12500	1485.0	121.0	16.37	1	0
4	1990	7000	1485.0	121.0	16.37	1	0

Area_Argentina	Area_Armenia	….	Item_ Cassava	Item_ Maize	Item_ Plantains and others	Item_Plantains and others
0	0	….	0	1	0	0
0	0	….	0	0	0	0
0	0	….	0	0	0	0
0	0	….	0	0	0	0
0	0	…	0	0	0	0

Item_ Potatoes	Item Rice, paddy	Item_ Sorghum	Item_ Soybeans	Item_ Sweet potatoes	Item_ Wheat	Item_ Yam
0	0	0	0	0	0	0
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0

Where, $\overline{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$; $S_k = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ik}-\overline{x}_k)^2}$.

The correlation coefficient covariance matrix of standardized indicators should be computed using the basic equation for determining Pearson's correlation coefficient. The correlation coefficient is calculated using equation (4):

$$r = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} = \frac{\sum (x-\overline{x})(y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2}\sqrt{\sum (y-\overline{y})^2}} \tag{4}$$

Using principal component analysis and the correlation coefficient matrix R, get the Eigen values, eigenvectors, variance contribution rate, and cumulative contribution rate for each principal component. Make an assessment model for the index; execute a full review and analysis by generating a comprehensive score based on the contribution rate and the evaluation model's score; and then perform analysis such as ranking.

3.4. Feature selection lasso regression

A linear regression method called LASSO regression makes use of regularization and variable selection strategies to increase model accuracy. The total of the observations, predictor variables, and intercepts are what it seeks to minimize. The selection of variables is the outcome of the L1 consequences term, which promotes sparsity in coefficient estimations. High-dimensional datasets and multicollinearity across predictor variables are advantageous applications of LASSO, particularly in the prediction of hybrid crop production. LASSO makes it easier to choose important characteristics by encouraging sparsity, which might result in a crop yield forecast model that is more efficient and accurate.The section explains how feature selection for agricultural production prediction is aided by the linear regression approach known as LASSO regression. By minimizing the total number of assessments, predictor variables, and intercepts, LASSO uses variable selection and regularization approaches to improve model accuracy.

$$K_{lasso}\beta = \sum_{j=1}^{m} z_j - \sum_{i}^{m} w_{ji}\beta_i + \lambda \sum_{i=1}^{o} \beta_i \tag{5}$$

Where z_j is a result, w_{ji} is a covariance, λ is a quantity of a reduction, and β is a factor of regression analysis.

3.5. Ensemble regression based extra tree regressor

In this approach, we utilize an average of all the model predictions to conclude. In regression problems or when computing probabilities for classification difficulties, averaging can be used to make predictions. To improve the reliability of the predicted outcomes of models, ensemble methods combine the results of several models rather than just one. The combined models result in a significant increase in the accuracy of the predictions. Involved in the agricultural yield forecast process, our suggested ensemble regression approach produced high prediction accuracy. The field of supervised machine learning that is used to forecast or estimate agricultural yields include ensemble regression. From the ongoing observations of the input variables (independent variables) and the output variables, the learning occurred during the supervised ML approach (target variables). The same is true for our crop yield forecast model; it depends on continuous learning from site-specific separate variables including soil, the climate, and features of agricultural operations that fluctuate annually. The target variables in this study will be the crop yield information. The model we have offered for forecasting crop yield falls inside the process of regression genre types because it is used in the agriculture yield forecast industry.Two of the most well-known types of supervised machine learning are regression and classification. Using a range of feature extraction techniques, we evaluated our suggested crop yield prediction models, and we contrasted the crop yield accuracy outcomes using a range of machine learning regression algorithms. This allowed us to understand better how accurate our models were. The ER crop yield prediction technique is represented by Algorithm 1.

Algorithm 1. Ensemble regression

Given: Collection of Occurrences $A = \{a_i \epsilon\ R^M\}$*.Collection of Estimations* $B = \{b_i \epsilon\ N\}$*.Collection of training information* $D = \{a_i,\ b_i\}$.

I/p*:* $D = \{(a_i, b_i)\}\ |a_i, \in A, b_i \in B\}$ +

O/p*: An Ensemble prediction P*

Phase 1: Take level predictions

For $t \leftarrow 1\ to\ T\ do$

Take a base prediction P_t *based on D*

Phase 2: Generate a new dataset from D

For $i \leftarrow 1\ to\ m\ do$

Construct a newly extracted data set containing $\{a^{new},\ b_i\}$ *where* $a_i^{new} = \{p_j(a_i) for\ j = 1\ toi\ T$

Phase 3: Take2nd level, predictions

For $t \leftarrow 1\ to\ T\ do$

Take a base prediction p_t *based on the D*

Generate a newly extracted data set containing $\{a_i^{new},\ b_i\}$ *where* $a_i^{new} = \{p_j(a_i) for\ j = 1\ toi\ T$

Phase 4: Take 3rd-level predictions

Take a new prediction P^{new} *according to the recently retrieved data.*

Return $P(x) = P^{new}(p_1(a), p_2(a) \ldots., p_T(x))$

Using the Extra-Tree approach is equivalent to using Extremely Randomized Trees. In the context of numerical input characteristics, the selection of an appropriate cut-point is what determines a significant amount of the variation in the induced tree. This is the primary goal of using a random method for creating trees.

In terms of bias, abandoning the bootstrapping notion provides a distinct benefit from a statistical point of view. In most cases, the cut-point randomization has a significant impact on the reduction of variation. This approach produces the best outcomes for a considerable number of high-dimensional and difficult situations. From a computational point of view, the extra-tree conduct produces piece-wise multilinear approximated, as opposed to the piece-wise constant approximations produced by the random forest (RF) method.

4. Result analysis

Through comparative and assessment of outcomes, the proposed ER-ETR approach is comprehensively assessed. The return on investment calculation aid provided by the aforementioned model demonstrates the value and efficiency of the suggested approach in comparison to alternative approaches such as XGBoost, Convolution Neural Network – Recurrent Neural Network (CNN-RNN), Convolution Neural Network – Long Short Term Memory (CNN-LSTM), Bernoulli Deep Belief Network (BDN), Bayesian Artificial Neural Networks (BAN), and Deep Reinforcement Learning (DRL).The graphic shows the exactitude rate applied by the recommended approach for a "Mean absolute error (MAE), Mean square error (MSE), Root mean square error (RMSE), and Mean Absolute Percentage Error (MAPE)." Feature optimization is the procedure of selecting which are most advantageous characteristics according to simulation and prediction outcomes.

- **Mean absolute error (MAE)**, a gauge for the discrepancies between paired observations describing every same phenomenon, is defined as the fundamental difference between the projected value and the actual value. A comparison of Mean Absolute Error is seen in Table 2. Equation (5) shows the mean absolute error.

$$MAE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}|\widehat{z}_j - z_j|} \tag{5}$$

Fig. 3 displays the suggested systemsMAE along with a statement pointing out the inaccuracy in the forecasted total mean consumption for the current systems and the suggested approach. The suggested

Table 2
Comparison of Mean fundamental error.

Existing Methods	Mean absolute error
XGBoost	0.0049
CNN-RNN	0.0062
CNN-LSTM	0.0081
ER-ETF [Proposed]	0.0085

Table 3
Comparison of MSE.

Existing methods	MSE
XGBoost	0.0095
CNN-RNN	0.0091
CNN-LSTM	0.0089
ER-ETF [Proposed]	0.0085

Table 4
Comparative analysis of RMSE.

Existing methods	Root Mean square error
XGBoost	0.093
CNN-RNN	0.097
CNN-LSTM	0.098
ER-ETF [Proposed]	0.092

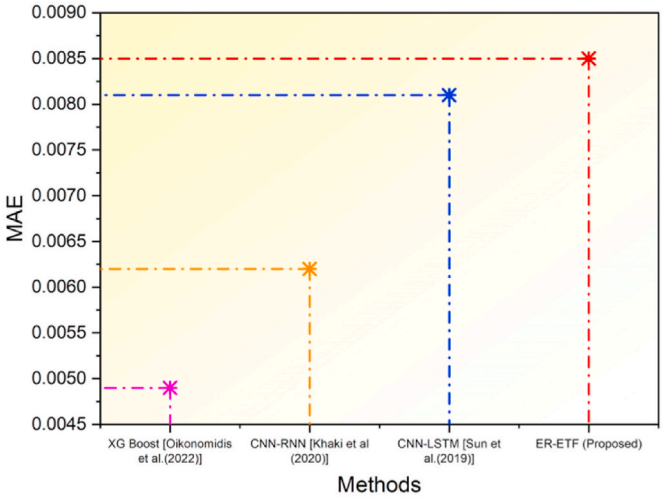

Fig. 3. Mean absolute error.

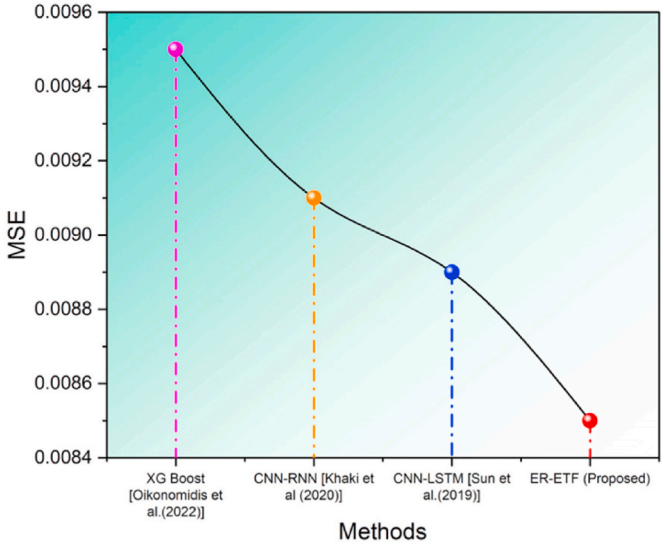

Fig. 4. Mean square error.

system has obtained 0.0085 of MAE, while CNN-LSTM has obtained 0.0081, CNN-RNN has achieved 0.0062, and XGBoost has obtained 0.0049. It implies that, in terms of value, the current strategy is better than the suggested one.

➢ **Mean Square Error** The total quadratic variability of the outcomes of an actual and identified variable multiplied by the total significance of the variable in questionis the mean square error, or MSE. The results of the comparison of mean square errors are presented in Table 3.Equation (6) shows the mean square error.

$$MSE = \frac{1}{m}\sum_{j=1}^{m}\left(z_j - \widehat{z}_j\right)^2 \tag{6}$$

The recommended system's mean square error and consumption estimates based on MSE for the proposed method and the existing systems are shown in Fig. 4. MSE-wise, CNN-RNN scores 0.0091, CNN-LSTM scores 0.0089, XGBoost scores 0.0095, and the recommended system scores 0.0085. Since the proposed method's value is less than the existing systems, it proves that our proposed model performs better than previous models.

➢ **Root mean square error (RMSE)** it is an evaluation of the statisticians' performance. The square root of the MSE yields the RMSE, which denotes the residuals' standard deviation. Table 4 displays the comparative study of RMSE. According to Equation (7), the RMSE.

$$MSE = \frac{1}{m}\sum_{j=1}^{m}\left(z_j - \widehat{z}_j\right)^2 \tag{7}$$

The RMSE of the suggested system is displayed in Fig. 5, along with a consumption projection for both the suggested technique and the current systems. The proposed system has gained 0.092 of RMSE. It shows that the proposed approach has a lower value than the existing models.

➢ **R square value (R^2)** is an indicator of statistics that expresses the variance that constitutes a model that uses regression. That can be attributed to independent variables. It has a 0–1 range, where 1 is the ideal fit. In predicting agricultural production, it evaluates the goodness of fit and predictive power of the model, putting a number on how well it accounts for crop output variability.

$$R^{2=1-}\frac{Sum\ of\ Squared\ Reciduals}{Total\ Sum\ of\ Squares} \tag{8}$$

The proposed system's R square is displayed in Fig. 6, along with a consumption forecast of the RMSE for both the existing and suggested systems. With CNN-RNN achieved 0.91, XGBoost reached 0.87, and

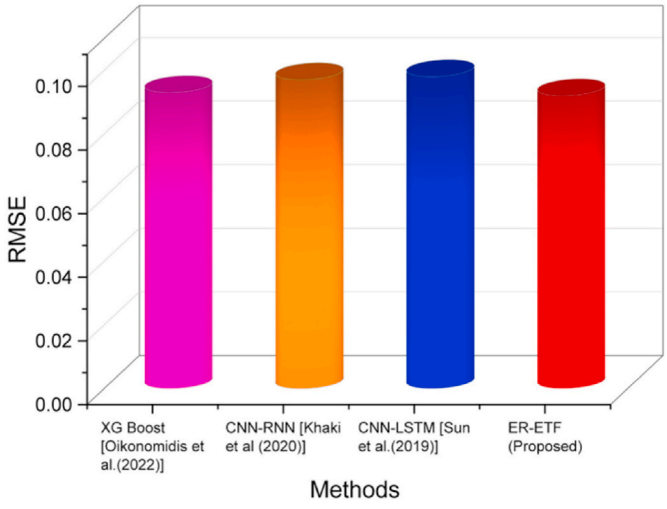

Fig. 5. Root mean square error.

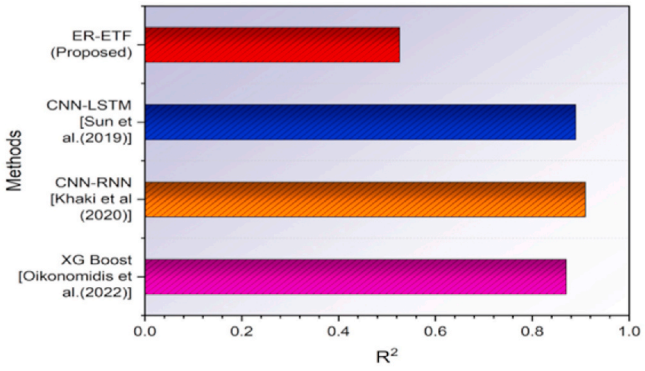

Fig. 6. Root mean square error.

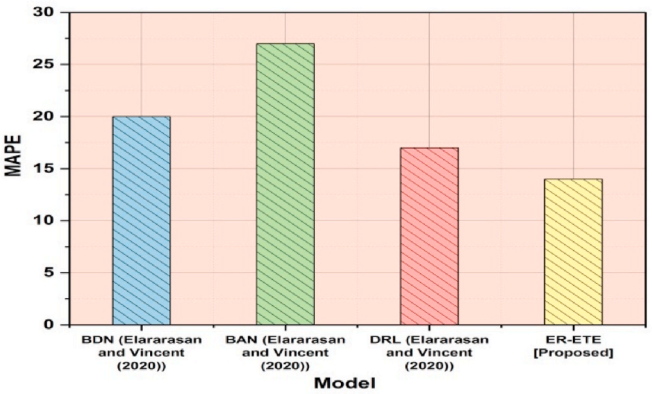

Fig. 7. Root mean square error.

CNN-LSTM reached 0.89 the suggested system achieves 0.526 of R square. It demonstrates that the proposed method is less valuable than the current models.

➢ **Mean Absolute Percentage Error (MAPE)** The ER-ETR is determined using the MAPE, which provides a thorough evaluation of predicting performance in agricultural output of prediction tasks. MAPE-wise, BDN scored 20, BAN scored 27, DRL scored 17, and the recommended system ER-ETF scores 14.Table 6 shows the comparison of MAPE. Fig. 7 shows the outcome of MAPE.

The evaluation metrics accuracy, precision, recall, f1-score are evaluated with current existing approaches such as Decision Tree [13], KNN [14], CNN-RNN (Varghese &Kandasamy 2021).

➢ **Accuracy (%)** The following formula is used to determine accuracy, which is defined as the number of forecasts totaled with the amended estimate.Equation (9) shows the measurement of accuracy.

$$Accuracy = \frac{Correct\ prediction}{Total\ number\ of\ prediction} \tag{9}$$

➢ **Precision (%)** is the proportion of accurately predicted beneficial findings to all expected positive observations. It assesses the degree of accuracy of the optimistic projections. The below formula is used to determine the accuracy.

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{10}$$

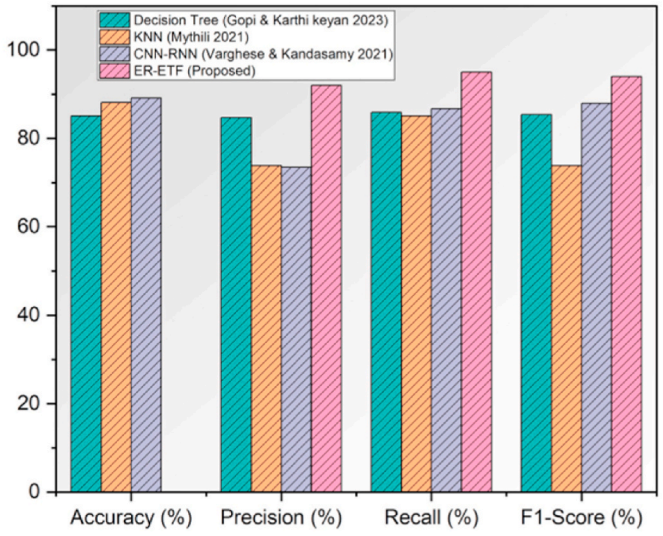

Fig. 8. Accuracy.

Table 5
Comparative analysis of R.^{2}.

Existing methods	R^2
XGBoost	0.87
CNN-RNN	0.91
CNN-LSTM	0.89
ER-ETF [Proposed]	0.526

Table 6
Comparative analysis of MAPE.

Model	MAPE
BDN	20
BAN	27
DRL	17
ER-ETF [Proposed]	14

Table 7
Comparative analysis of accuracy.

Evaluation Matrices	Decision Tree [13]	KNN [14]	CNN-RNN (Varghese &Kandasamy 2021)	ER-ETF (Proposed)
Accuracy (%)	85.07	88.14	89.16	95.00
Precision (%)	84.73	73.86	73.48	92.00
Recall (%)	85.91	85.09	86.73	95.00
F1-Score (%)	85.39	73.84	87.94	94.00

➢ **Recall (%)** The ratio of accurately anticipated favorable findings to genuine promising observation is known as recall. It assesses how well the model can account for every good example.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{11}$$

➢ **F1 Score (%)** is equal to the natural median of recall and accuracy. It provides an unbiased evaluation that considers both false positives and false negatives.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

Fig. 8 and Table 7 shows the Evaluation matrices of existing systems, and the proposed method is denoted. Its shows that the proposed approach is more effective than the existing Models (see Table 5).

5. Discussion

This study use contemporary methods to predict innovative Crop yield. They consist of Deep Learning (DL) and Machine Learning (ML). More DL algorithms were utilized, and the complexity increased, making DL algorithms challenging to manage and optimize. LSTM and RNN methods could have performed relatively better in terms of speed and Accuracy Sun et al. [15]. XGBoost, an algorithm that focuses on gradient boosting, is one of the most well-known methods for handling large and complex datasets with outstanding performance and speed Oikonomidis et al. [16].CNN-RNN models generated results that were nearly identical with less processing power, less complexity, and more specialized techniques Khaki et al. [17]. They are slower due to their capacity for managing time. Thus, the decision between using an ML or DL technique arises. A method's complexity only sometimes equates to its effectiveness.

6. Conclusion

In this work, we looked at hybrid models that included ML and DL approaches to estimate crop yield. To anticipate the crops, we suggest using a method called Ensemble Regression-based Extra Tree Regressor (ER-ETR), which combines deep learning and machine learning. The normalization procedure is used to gather and preprocess the dataset. If the suggested models are tested on various datasets, our findings can vary greatly. The testing criteria used to evaluate the classification model were MAE at 0.085, MSE at 0.085, RMSE at 0.092 and R^2 at 0.526. Each proposed typically outperforms competing models and also has a 95 % accuracy rate in predicting crop yield while preserving the original distribution of the data. Because farmers need to be informed, they will receive precise details on the ideal crop output on their portable electronics. With this, the task can be controlled at that exact moment without losing any money, even if the rancher is at home. Significant advancements in the agribusiness sector will further assist farmers in growing crops.

Code availability

Not applicable.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

There is no funding

References

[1] S. Agarwal, S. Tarar, A hybrid approach for crop yield prediction using machine learning and deep learning algorithms, J. Phys. Conf. 1714 (1) (2021) 012012. IOP Publishing.
[2] N. Kim, K.-J. Ha, N.-W. Park, J. Cho, S. Hong, Y.-W. Lee, A comparison between major artificial intelligence models for crop yield prediction: case study of the midwestern United States, 2006–2015, ISPRS Int. J. Geo-Inf. 8 (5) (2019) 240.
[3] X. Xu, P. Gao, X. Zhu, W. Guo, J. Ding, C. Li, M. Zhu, X. Wu, Design of an integrated climatic assessment indicator (ICAI) for wheat production: a case study in Jiangsu Province, China, Ecol. Indicat. 101 (2019) 943–953.
[4] D. Batool, M. Shahbaz, H. Shahzad Asif, K. Shaukat, T.M. Alam, I.A. Hameed, Z. Ramzan, A. Waheed, H. Aljuaid, S. Luo, A hybrid approach to tea crop yield prediction using simulation models and machine learning, Plants 11 (15) (2022) 1925.
[5] J. Cao, H. Wang, J. Li, Q. Tian, D. Niyogi, Improving the forecasting of winter wheat yields in northern China with machine learning–dynamical hybrid subseasonal-to-seasonal ensemble prediction, Rem. Sens. 14 (7) (2022) 1707.
[6] S. Khaki, L. Wang, Crop yield prediction using deep neural networks, Front. Plant Sci. 10 (2019) 621.
[7] L.O. Colombo-Mendoza, M.A. Paredes-Valverde, M.D.P. Salas-Zárate, R. Valencia-García, Internet of Things-driven data mining for intelligent crop production prediction in the peasant farming domain, Appl. Sci. 12 (4) (2022) 1940.
[8] P. Feng, B. Wang, D. Li Liu, C. Waters, D. Xiao, L. Shi, Q. Yu, Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning techniques, Agric. For. Meteorol. 285 (2020) 107922.
[9] M. Shahhosseini, G. Hu, I. Huber, S.V. Archontoulis, Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt, Sci. Rep. 11 (1) (2021) 1–15.
[10] M. Cooper, T. Tang, C. Gho, T. Hart, G. Hammer, C. Messina, Integrating genetic gain and gap analysis to predict improvements in crop productivity, Crop Sci. 60 (2) (2020) 582–604.
[11] K.S.M. Anbananthen, S. Subbiah, D. Chelliah, P. Sivakumar, V. Somasundaram, K. H. Velshankar, M.A. Khan, An intelligent decision support system for crop yield prediction using hybrid machine learning algorithms, F1000Research 10 (2021).
[12] M. Keerthana, K.J.M. Meghana, S. Pravallika, M. Kavitha, An ensemble algorithm for crop yield prediction, in: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE, 2021, February, pp. 963–970.
[13] S.R. Gopi, M. Karthikeyan, Effectiveness of crop recommendation and yield prediction using hybrid moth flame optimization with machine learning, Eng. Technol. Appl. Sci. Res. 13 (4) (2023) 11360–11365.
[14] K. Mythili, A swarm-based bi-directional LSTM-enhanced Elman recurrent neural network algorithm for better crop yield in precision agriculture, Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12 (10) (2021) 7497–7510.
[15] J. Sun, L. Di, Z. Sun, Y. Shen, Z. Lai, County-level soybean yield prediction using deep CNN-LSTM model, Sensors 19 (20) (2019) 4363.
[16] A. Oikonomidis, C. Catal, A. Kassahun, Hybrid deep learning-based models for crop yield prediction, Appl. Artif. Intell. (2022) 1–18.
[17] S. Khaki, L. Wang, S.V. Archontoulis, A cnn-rnn framework for crop yield prediction, Front. Plant Sci. 10 (2020) 1750.