sciendo

RURAL SUSTAINABILITY RESEARCH
52(347), 2024
ISSN – 2256-0939

Received: 20 February 2024 Revised: 13 August 2024 Accepted: 12 November 2024

Comparative Analysis on Crop Yield Forecasting using Machine Learning Techniques

Shubham Sharma[1], Gurleen Kaur Walia[2], *Kanwalpreet Singh[1,3], Vanshika Batra[1], Amandeep Kaur Sekhon[1], Aniket Kumar[1], Kirti Rawal[2], Deepika Ghai[2]

[1] Department of Computer Science and Engineering (UIET, Hoshiarpur), Panjab University, Chandigarh, India
[2] School of Electronics and Electrical Engineering, Lovely Professional University, Phagwara, Punjab, India
[3] Chandigarh University, Mohali, India

Abstract. Global overpopulation necessitates increased crop yields, yet available arable land is limited. The study compares and evaluates the performance of three machine learning algorithms—Random Forest (RF), Extra Trees (ET), and Artificial Neural Network (ANN)—in crop yield prediction. Using 28,242 samples with seven features from 101 countries, we evaluated these models based on Mean Absolute Error (MAE), R-squared (R^2), and Mean Squared Error (MSE). The ET regression model demonstrated superior performance, achieving an MAE of 5249.03, the lowest among the models tested. Despite having the highest R^2 value of 0.9873, the ANN exhibited higher MAE and MSE values, indicating less reliability. The RF model showed intermediate results. With a prediction accuracy of 97.5%, the ET model proved to be the most effective for crop yield prediction, achieving the highest accuracy reported to date. Future research should explore more advanced algorithms and larger datasets to validate these findings further.

Key words: crop yield, machine learning, deep learning, regression, prediction, extra trees, random forest, artificial neural network.

Introduction

Agriculture is one of the fields which leads to the growth of the economy. Around 50% of the world's population is involved in agricultural activities such as growing crops, fruits, vegetables, flowers and nurturing of livestock. With the increasing population, the demand for food is increasing day by day but the area of sowing is limited. Accurate crop yield forecasting plays a crucial role in optimizing agricultural practices, allowing farmers and policymakers to make informed decisions about resource allocation, planting strategies, and distribution logistics, ultimately helping to ensure food security and meet increasing global food demands (Fritz *et al*., 2019). The crop yield is highly affected by various challenges like area of cultivation, level of irrigation, climatic conditions, genotype, and many others, making crop yield prediction difficult. The data provided during the training phase is another aspect that affects the prediction in addition to the number of parameters. To serve this purpose, a lot of studies have been done using different machine learning and related fields. The accurate pre-information regarding the crop yield leads to loss minimization. Machine learning is the fast-growing approach that helps in the forecast of crop yield. The main idea of using machine learning models is that the agriculture sector should be able to increase its productivity. Precision agriculture would be the focus, in which desirable environmental factors would be weighed against quality assurance. Forecasting of crop yield helps the farmer to know beforehand what to grow and when to grow.

In this research, we address the crucial task of crop yield forecasting using regression algorithms, namely Random Forest (RF), Extra Trees (ET), and Artificial Neural Networks (ANNs), which are widely recognized for their effectiveness in predictive modeling. Random Forest (RF) is an ensemble machine learning algorithm that uses Bootstrap Aggregation (Bagging) to improve predictive accuracy. During training, multiple decision trees are generated as base learning models, with the final prediction being the average of the predictions from each decision tree. The independent,

* Corresponding Author's email: malhi.kanwalpreet@gmail.com

DOI: 10.2478/plua-2024-0015

parallel nature of decision trees in RF results in robust predictions with lower variance. The Extra Trees (ET) model, similar to RF, also creates decision trees but introduces randomness by selecting feature splits and samples randomly. This further reduces correlation between the trees, increasing performance. The model is highly efficient and performs well for complex datasets. Finally, Artificial Neural Networks (ANNs) are computational models inspired by biological neural networks. ANNs consist of interconnected nodes (neurons) in layers: input, hidden, and output layers. Their capacity to model nonlinear relationships makes them particularly effective in solving complex problems, such as crop yield prediction, where they can handle high-dimensional data and process it in parallel. Our primary objective is to assess the performance of these algorithms in predicting crop yields based on pre-processed datasets. Specifically, we focus on evaluating the Mean Squared Error (MSE), R-squared (R^2), and Mean Absolute Error (MAE) metrics to compare the predictive accuracy of the three techniques.

Our primary contribution lies in the comparative evaluation and validation of three distinct machine learning models—RF Regression, ET Regression, and Artificial Neural Network—tailored specifically for crop yield prediction in a structured agricultural dataset. By systematically analyzing each model's performance, our work identifies the ET Regression model as the optimal solution, balancing accuracy with computational efficiency. Furthermore, our research introduces a practical approach that can aid farmers and policymakers by predicting crop yields with high precision, offering valuable insights on optimal planting schedules and resource allocation. Unlike prior studies, which often lack such focused comparisons, our work presents a robust framework that integrates accurate yield forecasting with actionable recommendations, setting the foundation for more informed agricultural decisions. This framework supports sustainable practices by helping farmers anticipate crop yield outcomes based on specific inputs, enabling better planning in alignment with ecological and economic needs.

This study builds upon the extensive research in the field, drawing comparisons with prior works. For instance, (Nigam *et al*., 2019) utilized an RF regressor alongside other algorithms, achieving a maximum accuracy of 64.80%. P.S. Maya Gopal, 2019 employed the ANN model, reporting Root Mean Squared Error (RMSE) of 5.1% and MAE of 6.4%. (Kim *et al*., 2019) conducted a comparative analysis involving RF, ANN, and ERT models, revealing MAE values of 0.708, 0.705, and 0.703 and RMSE values of 0.929, 0.928, and 0.922 for RF, ANN, and ERT, respectively. (Mishra *et al*., 2016) achieved 96% accuracy using the ANN model. Additionally, (Sreerama & Sagar, 2020) preferred other regression models over RF regressor due to comparatively lower R^2 scores. (Dahikar & Rode, 2014) conducted an analysis using ANN on various crops, affirming its efficacy in predicting crop yields. (Kuwata & Shibasaki, 2016) highlighted the usefulness of ANN in extracting significant features from high-dimensional data for crop yield studies. (Feng *et al*., 2020) combined an RF regression model with growth stage-specific markers to develop a hybrid yield forecasting technique for wheat.

This study investigates reliable regression techniques such as RF, ET, and ANNs for forecasting crop yields, supported by contemporary methodologies and data pre-processing methods, ensuring the study's reliability and alignment with current agricultural research. Through our analysis, we aim to contribute valuable insights to the field, informing future research and practical applications in agricultural forecasting.

Related works

Numerous studies highlight how agricultural productivity is impacted by environmental variables (Swain *et al*., 2024) such as temperature, rainfall, and pesticide. This is consistent with our research of the Kaggle dataset, which revealed a substantial association between these parameters and yield (as indicated by an R-squared value of 0.98) along with crop type (with an emphasis on potatoes). This shows that crop yields may be accurately predicted by machine learning algorithms by capturing these correlations.

This study looks into using machine learning to predict crop yields in Rajasthan, India (Jhajharia *et al*., 2023). Drawing from prior studies that emphasize the impact of variables such as temperature, rainfall, and soil composition on crop productivity, we investigate the efficacy of multiple algorithms, such as Random Forest, SVM, Gradient Descent, Long Short-Term Memory (LSTM), and Lasso Regression. Our results support earlier research that suggested Random Forest performed well (obtaining an R-squared of 0.963), but they also show that larger datasets may be needed for deep learning models like LSTM to function at their best. To summarize, this study not only highlights the promise of machine learning to forecast crop output, but also notes that more research into environmental factors and the development of several new deep learning models with more accessible data are needed.

Numerous works have been done on crop yield prediction using machine-learning approaches such as ANN, LSTM, CNN, etc. (Bodapati *et al*., 2022) grouped the data using the ANN technique during the training phase, and the yield produced was estimated

during the testing phase. The model took into account variables such as soil, weather, and moisture content, because of which it could not provide accurate findings. By considering more factors, such as temperature, soil quality, seed quality, and a wide range of others, as well as by employing a variety of algorithms, forecasts could have been made with more accuracy. ANN was found to be the best technology available for correctly predicting crop yields. The accuracy it attained was 80%, which is the highest of all other algorithms used. (Pandith *et al*., 2020) also used ANN for predicting mustard yield and gave 76.86% accuracy with 99.61% precision.

Three classifier models (Venugopal *et al*., 2021) namely Logistic Regression, Naïve Bayes (Kumar et al., 2020), and RF (Jeong *et al*., 2016) were used. Their system of predicting crop yield included manual counting, climate-smart pest management, and satellite imagery, which made the results unreliable. Many factors affect crop yield and production. Among all these factors, the research considered temperature, rainfall, area, humidity, and wind speed as the dominating ones. In comparison to the other two algorithms, RF appeared to be the most accurate of the three used. In RF, the data was trained using the bagging method, which increased accuracy and results by 92.81%. The work (Suresh *et al*., 2021) employed a dataset that took into consideration factors such as soil characteristics, soil type, pH value, climate parameters, wind, rainfall, humidity, temperature, cultivation costs, and manufacture. The first phases were data gathering and cleansing, then connecting to modules like weather and temperature forecast. The yield was then forecasted once the model was trained using the RF Algorithm and filled with input parameters like district name, crop name, area, and type of soil. The study demonstrated the usefulness of data mining approaches for forecasting agricultural production based on input variables and climatic conditions. The other grains such as jowar, potato, etc. and districts of Maharashtra like Amravati considered in the analysis had a reliability of prediction above 75%, suggesting higher predictive capability.

The dataset based on 32 districts of Tamil Nadu was taken from the Tamil Nadu Agricultural University (TNAU) website for forecasting soil (Chandraprabha & Dhanaraj, 2018) which helped in crop yield prediction. It described that soil is the key component. To determine crop yield, nutrients and pH values of the soil were considered. To predict the type of crop suitable for the soil, algorithms such as Naïve Bayes, Bayes Net, and Instance-based K-Nearest Neighbor (IBK) were used to consider the total production and area sown district-by-district. Matthew's Correlation Coefficient (MCC), precision, recall, f-measure, and true positive value were used to determine accuracy. It was found that RF shows a maximum level of accuracy of 97.4%. However, IBK, Naïve Bayes, and Bayes Net showed accuracies of 97.31%, 53.61%, and 68.81%, respectively. It was also concluded that the RF algorithm works well in the small dataset. (Elavarasan & Vincent, 2021) used a hybrid reinforcement RF algorithm that helped in crop yield prediction. This helped in selecting the variables to split and effectively train data. In the research (Oikonomidis, Catal, & Kassahun, 2022), XGBoost machine learning, Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), CNN-XGBoost, recurrent neural network (RNN), and CNN-LSTM were used out of which the hybrid CNN-DNN model outperformed other models with an RMSE value of 0.266, an MSE of

0.071 and an MAE of 0.199. In the work (Khaki & Wang, 2019), crop yield detection was implemented using DNN as the main technique with 81.44% accuracy.

In (D. Patel & G. Patel, 2021), various machine learning algorithms are used including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), RF, and ANN. These algorithms predicted the crop suitable for different seasons and soil parameters. From all these above-mentioned algorithms, KNN was observed to have the highest obtained accuracy of 95.2%. The work (Iniyan, Varma, & Naidu, 2022) considered the district-level data of Maharashtra. It performed multiple linear regression, ridge regression, partial least squares regression, lasso regression, gradient boosting regression, LSTM, and Elastic-Net Regression and found a maximum accuracy of 86.3% using LSTM. Using the dataset of 13 US states from 2016 to 2018 to predict soybean and corn yields, the precision of 77.84 and 75.04 were obtained, respectively, by using the CNN-RNN model (Khaki, Wang, & Archontoulis, 2020). A hybrid approach-machine learning models and simulation was used (Batool *et al*., 2022) to predict the tea yield. It got the values of MAE of 0.123 tonnes per hectare (t/ha), RMSE of 0.154 t/ha, and MSE of 0.024 t/ha using the XGBoost regressor. It resulted that the machine learning regression algorithm performed better in yield prediction using less data than the simulation model. The research (Satir & Berberoglu, 2016) estimated the crop yields of the three major crops, wheat, corn, and cotton, using Stepwise Linear Regression (SLR) and vegetation indices. It gave an accuracy of 61%, 46%, and 65% for wheat, corn, and cotton, respectively.

Considering technological advancements, the research (Nithya, Josephine, & Jeyabalaraja, 2023) employed a synthesis of the Internet of Things (IoT) and ML methodologies to prognosticate crop yields

within the Indian subcontinent. It collected the data using IoT-based sensors from north and south India and gave commendable accuracies of 97% and 96% using algorithms like KNN, Support Vector Machine, and logistic regression. The research (Dharwadkar, Kalmani, & Thapa, 2023) embarked on the development of a hybrid model, integrating LSTM and one dimensional (ID) CNN with an attention layer. The work was concentrated on the dataset having the predominant crops of India – rice and wheat. A comparative analysis was conducted, juxtaposing the hybrid model against other ML methods like RF Regressor, and Decision Tree Regressor, and found that the hybrid model had unparalleled performance with 0.967 R^2 value.

From the literature review, it is concluded that a substantial work focused on crop yield prediction using various machine learning techniques. The studies explored different machine and deep learning algorithms such as ANN, LSTM, CNN, RF, SVM, KNN, Multiple Linear Regression, Ridge Regression, Partial Least Squares Regression, Lasso Regression, Gradient Boosting Regression, Elastic-Net Regression, and hybrid models. The findings identified ANN as the best technology for accurately predicting crop yields, achieving an accuracy of 80%. Additionally, RF demonstrated promising results with an accuracy of 92.81% and was found to be the most accurate among the classifiers used. Furthermore, the integration of data mining approaches and climate parameters proved useful in forecasting agricultural production based on input variables and climatic conditions. The significance of soil characteristics, including nutrients, pH values, and suitability for specific crops, was emphasized in several studies. The use of vegetation indices and stepwise linear regression contributed to estimating crop yields for wheat, corn, and cotton, achieving accuracies ranging from 80% - 96%.

The literature review has identified several shortcomings in the results of the studies on crop yield prediction using machine learning approaches. One major limitation is the limited consideration of factors in some models, which can lead to potential inaccuracies in the predictions. By not incorporating a comprehensive range of variables such as temperature, pesticides, rainfall, and others, the models may fail to capture the full complexity of the crop yield dynamics. Additionally, reliance on manual counting, climate-smart pest management, and satellite imagery introduces uncertainties and can make the results unreliable. Certain algorithms, such as Naïve Bayes and Bayes Net, exhibited lower accuracy rates compared to other models, indicating their limitations in accurately predicting crop yields. Furthermore, some studies encountered challenges with small datasets, which can affect the robustness and generalizability of the predictions. In conclusion, these studies highlight the efficacy of machine learning techniques in crop yield prediction. The integration of various factors, such as soil characteristics, climate parameters, and extensive datasets, along with appropriate algorithm selection, plays a crucial role in achieving higher predictive capabilities. Future research should continue to explore novel approaches and consider additional factors to further enhance the accuracy and reliability of crop yield predictions.

Methodology

In this paper, judgments about managing agricultural risks are generated and projections are made using three models: RF Regressor, ET Regressor, and ANN. To identify the most effective machine learning algorithms for crop yield prediction, we employed the LazyRegressor class from the LazyPredict library, which facilitated a comprehensive evaluation of various regression models based on key performance metrics: R-Squared, RMSE (Root Mean Squared Error), and Time Taken for training. Among the models assessed, RF, ET, and ANN emerged as the top performers. RF and ET exhibited high R-Squared values, indicating strong predictive accuracy, while their RMSE scores were comparatively low, suggesting precise yield estimations. Additionally, both models demonstrated efficient training times, making them suitable for practical applications in agricultural forecasting. The first two models are regression models that examine the dependency of independent variables on the dependent variables (Seldon, 2021), whereas ANN is a deep learning technique that is developed from the concept of Biological Neural Networks (Singh, 2021). This method is tested on Apple M1 8GB Macintosh HD macOS Ventura 13.3.1(a) using Python. The flowchart of the proposed methodology for crop yield prediction is shown in Figure 1.

Data collection

Understanding global agricultural output requires the correct knowledge of crop yield history, weather factors (such as, temperature, rain, etc.), and pesticide usage. For the same reason, the publicly accessible datasets are provided by the World Data Bank, and Food and Agriculture Organization (FAO). The dataset (Crop Yield Prediction dataset) used for this research work contains 28242 samples with 7 features (R. Patel n.d.). It gives us the following informational data:

- *Areas:* The dataset covers 101 countries all over the world.
- *Items:* The dataset uses the numerical information of 10 crops that includes ‘Rice/ Paddy’, ‘Maize’, ‘Potatoes’, ‘Sorghum’,

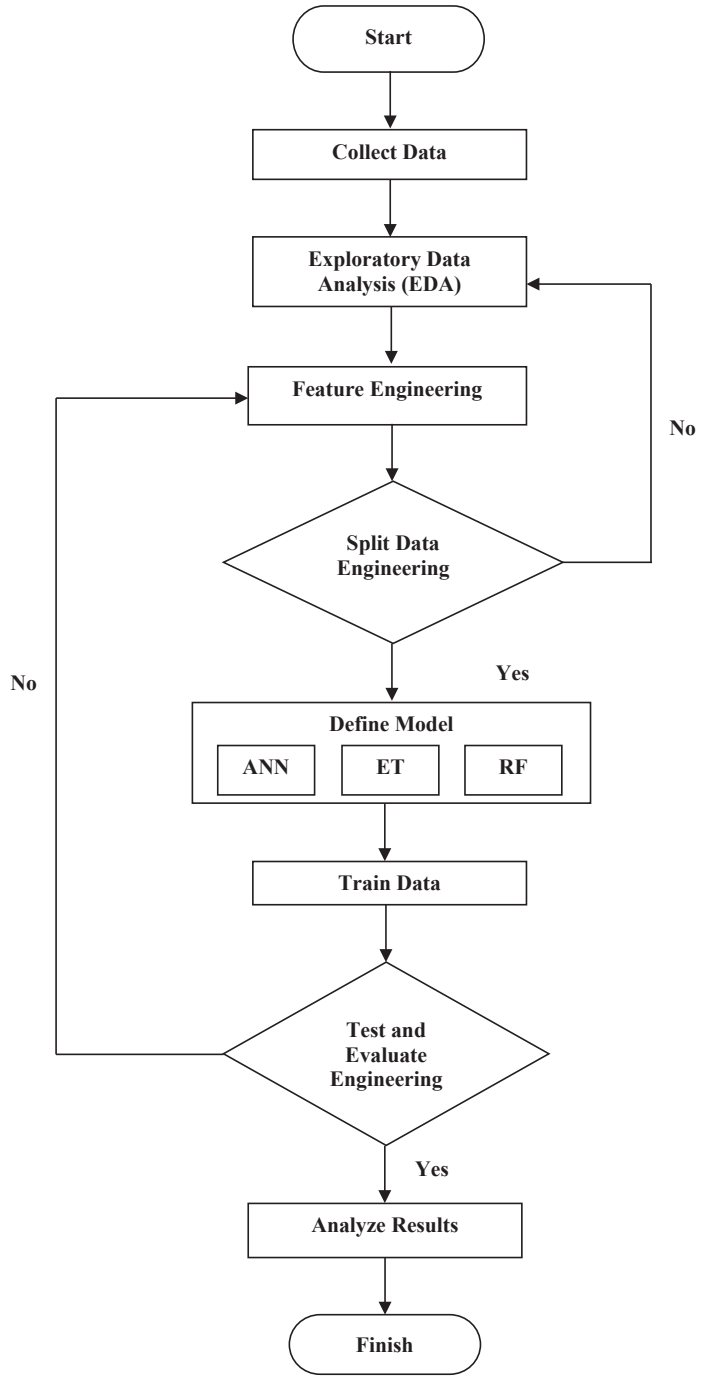


Figure 1. Flowchart of the proposed methodology for crop yield prediction.

'Soybeans', 'Wheat', 'Cassava', 'Sweet Potatoes', 'Plantains and others, and 'Yams'.

- *Years:* The dataset includes information from 1990 to 2016.
- *hg/ha_yield (hectogram/hectare yield):* It gives the crop yield production value.
- *Average Rainfall*: This paper utilizes rainfall per year information, gathered from World Data Bank from 1985 to 2017.
- *Pesticides*: The dataset utilizes data for pesticides from FAO that includes information about tonnes of pesticides used.
- *Average Temperature*: This data frame includes data about the country, year, and average recorded temperature from the year 1743 to 2013.

Data pre-processing or Exploratory Data Analysis (EDA)

At the beginning of the analytical process, the initial analysis of the dataset has been carried out which is termed EDA (Kanyutu, 2023). As an outcome, the outlier existence and imbalance, if any, could be detected. EDA is the process of looking at or comprehending the data and drawing conclusions from the data. The elimination of duplicate values, treatment of dataset outliers and missing values, normalization and scaling, and analysis of the data distribution are all made easier with its assistance. Using EDA, the analysis of data can be done, either *numerically* or *categorically*, resulting in its visualization or statistical decision. In this paper, main emphasis is laid on the numerical data analysis as the numerical values (integers and float data- type values) are dealt with.

Categorical data analysis

Categorical data refers to the discrete data having data-type 'object', where the observations are organized into mutually exclusive ordered or unordered categories; such as name, job, company name, gender, address, etc. Here, we have categorical data under the names of 'item' (Varieties of Crops) and 'area' (Country Names).

To analyze categorical data, we considered *count plots* as the best option as it represents the occurrence of response variables similar to the bar graph. As the categorical data is nominal, we implemented the bar chart that displays the frequency of *items* and *areas.* Here, Figure 2 shows the count plots for the 10 crops (Maize, Potatoes, Rice (Paddy), Sorghum, Soybeans, Wheat, Cassava, Sweet Potatoes, Plantains and others and Yams) that we have considered, and Figure 3 shows the area-wise count plots of different countries.

Climatic conditions have a significant impact on the crop yield. Maize, for instance, suffers from reduced yields under high temperatures and drought conditions. Potatoes are vulnerable to frost and excessive rainfall, which can damage their tubers. Rice paddy requires ample water, making both drought and flooding problematic for its growth. Sorghum is relatively heat-tolerant and drought-resistant but still needs some moisture to thrive. Soybeans are sensitive to extreme temperatures and drought, which can negatively affect their yield. Wheat faces challenges from heat stress and frost, and it requires consistent moisture to grow effectively. Cassava thrives in warm conditions but is frost-sensitive and needs consistent rainfall. Sweet potatoes prefer warm temperatures but are susceptible to frost and excessive water, which can harm their yields. Plantains and similar crops need stable warmth, with frost and drought impacting their fruit development. Yams, too, prefer warmth but are sensitive to both frost and drought conditions. The analysis of Figure 2 and Figure 3 reveals key insights from two horizontal bar charts. In Figure 2, the chart displays the count of different types of crops, with the X-axis ranging from 0 to 4000 and the Y-axis listing crops such as Maize, Potatoes, Rice (paddy), Sorghum, Soybeans, Wheat, Cassava, Sweet potatoes, Plantains, and Yams. The length of each bar represents the quantity of each crop, with Potatoes having the

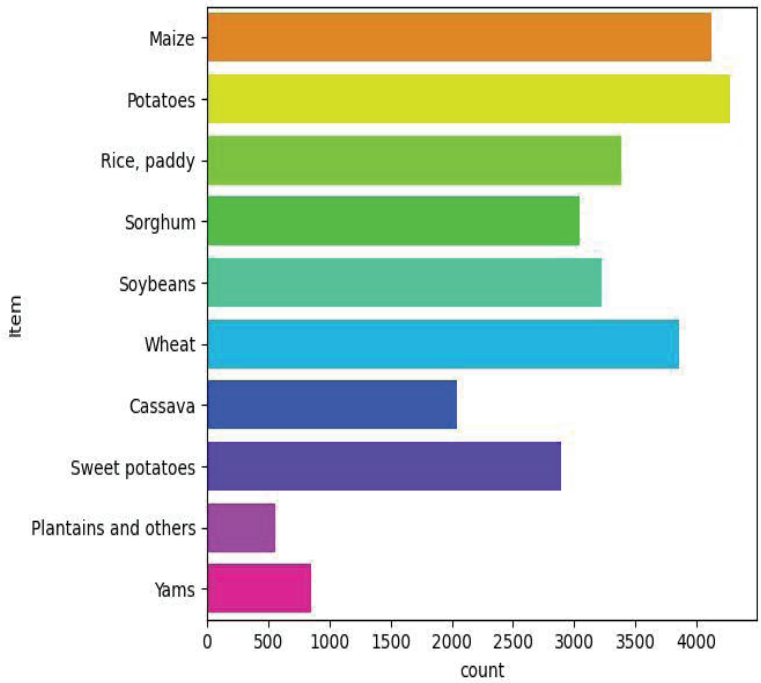


Figure 2. Count plots for items (variety of crops).

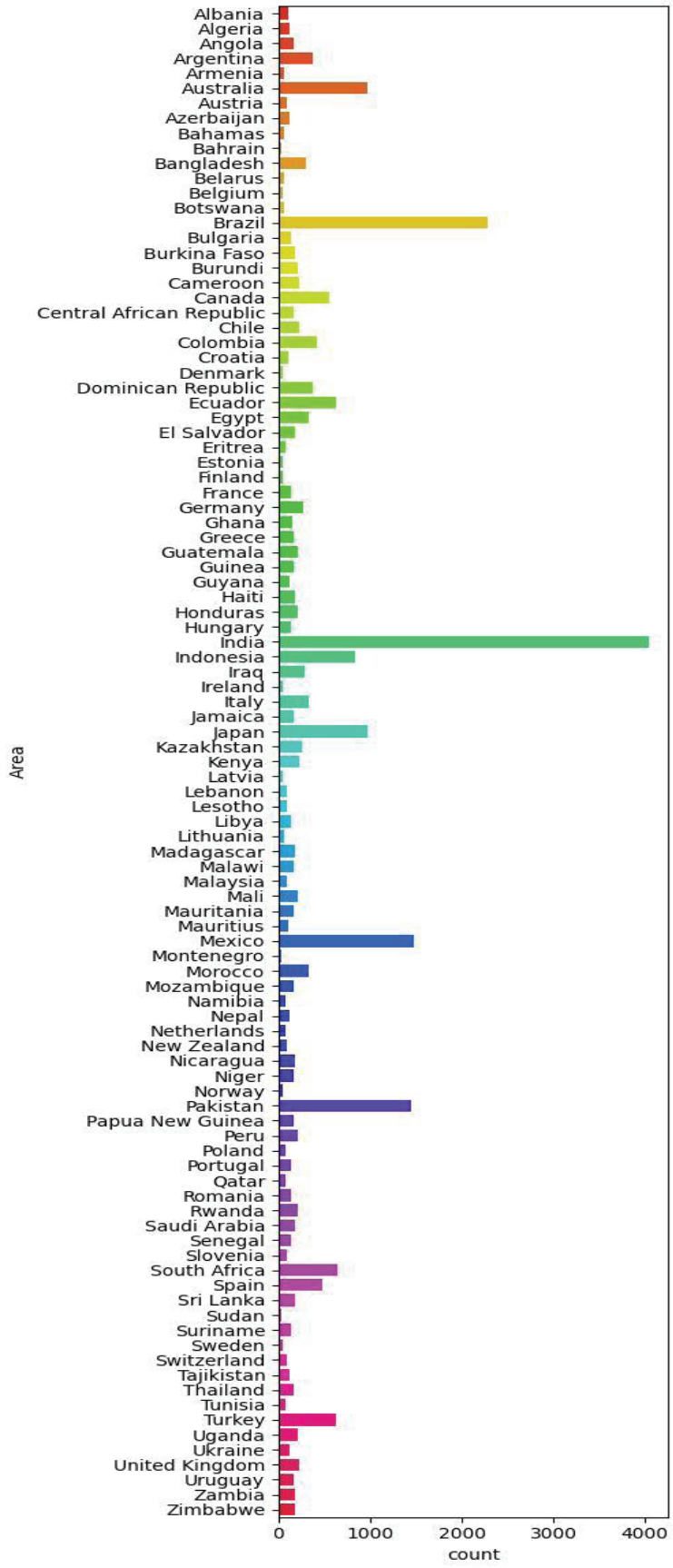


Figure 3. Count plots for Global Distribution of Crop Cultivation by Area.

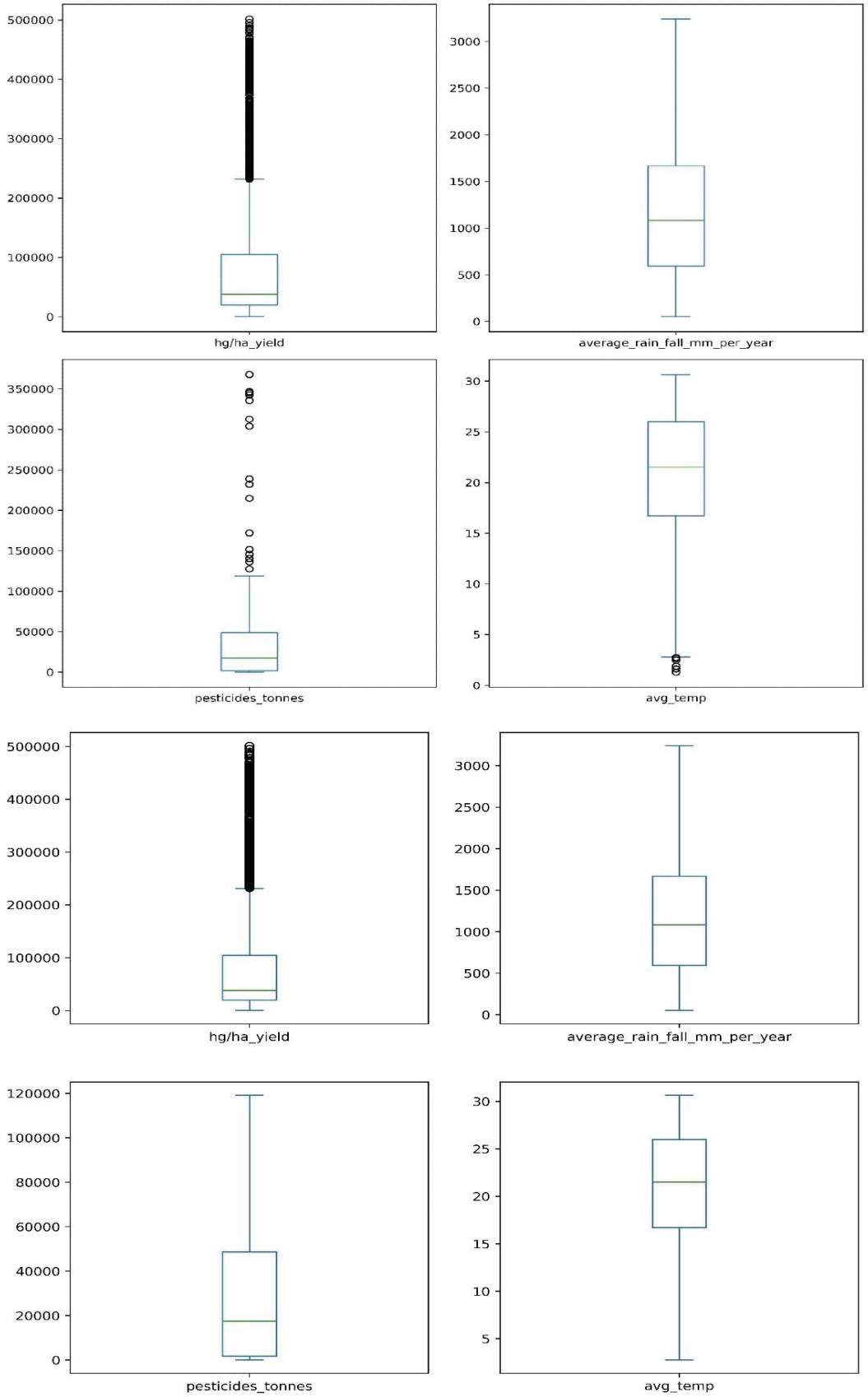


Figure 4. Box plot for numerical data before removing outliers and after removing outliers.

highest count, followed by Maize and Wheat. In Figure 3, the chart illustrates the count of available data points for crop yield prediction across various countries. The X-axis also ranges from 0 to 4000, and the Y-axis lists the names of different countries. The length of each bar indicates the number of data points available for crop yield prediction in that country, with India having the highest number of data points based on the data presented.

Numerical data analysis

Numerical data refers to the quantitative data with the data types of 'integers', 'float', etc., where the response variables are recorded in the numerical or quantitative form; such as salary, age, id, etc.

In this work, we have numerical data for the following features: 'year', 'hg/ha_yield', 'average_rain_fall_mm_per_year', 'pesticides_tonnes', and 'avg_temp'. To analyze the numerical data, *box plots* have been used. Box Plot is the summarized representation of 5 values evaluated from the General Statistical Analysis – min (minimum), 25% (lower quartile), 50% (median), 75% (upper quartile), and max (maximum).

The Figure 4 shows merged box plots for each variable, with the top row displaying plots with outliers and the bottom row displaying plots without outliers for pesticides_tonnes and avg_temp. For average_rain_fall_mm_per_year, the box plot remains unchanged, as this feature does not contain any outliers, maintaining a consistent distribution across the dataset.

The plot of pesticides_tonnes, hg/ha_yield, and avg_temp contain outliers, and the rest do not. These can be detected by seeing the data points that fall 1.5 times Inter-Quartile Range (IQR) above the 3rd quartile and below the 1st quartile. These are data points that deviate noticeably from the remainder of the dataset. As a result of poor data input or incorrect observations, they are frequently atypical observations that distort the data distribution. Thus, these must be reduced to get accurate results. However, we do not remove the outlier in the hg/ha_yield column because it can provide valuable insights into extreme or unusual yield conditions, which can be important for understanding and addressing potential factors affecting crop productivity. Moreover, the hg/ha_yield feature is treated as the output variable because it represents the primary outcome of interest in crop yield prediction.

Treating Categorical Value: Categorical variables can be transformed into dummy variables using a variety of techniques. It is a vital step in the pre-processing of data, which is a crucial component of machine learning or statistical models. After creating a bogus variable, a total of 114 features are obtained. To treat categorical values, the get_dummies method of the pandas Python library is used.

Some important results using EDA or data pre-processing:

- The outliers (an observation that lies an abnormal distance away from other values) are present in *hg/ha_yield, pesticides_tones,* and *avg_temp*, detected through Numerical Data Analysis, specifically, box plots.
- The dataset is bivariant as we have two features – *item* and *area* because they are of the 'object' data type.
- The *areas, items, hg/ha_yield, avg_temp, average_rain_fall_mm_per_year,* and *pesticides_tonnes* features of the dataset contain non-missing values.
- There is no imbalance in the *areas, items, hg/ha_yield, avg_temp, average_rain_fall_mm_per_year,* and *pesticides_tonnes* features of the dataset.

Feature engineering

The implementation of feature engineering on any model increases its chances to perform better. The performance of any model depends on two factors – data pre-processing and manipulation. Feature engineering involves nothing but the manipulation of the dataset to improve the accuracy of the model.

Feature selection

A good model is defined from the inter-dependencies among the features of the dataset and in the feature selection, independent features are selected that play a vital role in getting the desired outcome or the dependent features. Three categories are used, viz – Average rainfall (mm/year), areas, and items (crops). All the individual entries of the three categories contributed almost equally, therefore, no value is discarded (Ghai *et al*., 2022).

Handling missing values

During the data pre-processing stage, a thorough inspection of the dataset was performed to check for any missing values. After this inspection, we confirmed that the dataset was complete, with no missing values detected in any of the features. As a result, there was no need to apply imputation techniques or remove rows with missing data. This ensured that our data was ready for modeling without requiring additional handling for missing values.

Handling outliers

Outliers are the response variables that have a huge difference from other observations in the dataset that might be caused due to some error or the variability of the feature values. Outliers correspond to the skewness in the dataset. From the EDA – Numerical Data Analysis, we observed that the outliers are present in some of the features of the

dataset that might reduce the accuracy of our model. To ensure the higher performance of the model, it is highly recommended to handle the outliers in the most efficient way possible. According to (Bonthu, 2021), to handle outliers, there are three methods:

- *Trimming* involves deleting the outliers and copying the rest of the array as it is.
- *Capping and Flooring* involve setting the maximum quantile (capping) limit of the 90th percentile and minimum quantile (flooring) 10th percentile value. Outliers larger than the capping value are replaced with the 90th percentile value and outliers smaller than the flooring value are replaced with the 10th percentile value.
- *Mean/Median Imputation* involves replacing the outliers with the median value.

Feature scaling

It is considered that in some datasets, we might have the value with different units. And, the algorithm tends to weigh values according to their numerical values without considering their units. Here comes *feature scaling*, the technique to rescale the features, thus reducing the scattering effect (Huilgol, 2020). There are two techniques for feature scaling, one is *MinMaxScaler()* and another is *StandardScaler(). Min-max scaling* rescales the feature values between 0 and 1 whereas standard scaling rescales the data to have a mean of 0 and a standard deviation of 1.

In this paper, the *Min-Max Scaling* technique is used, using the *MinMaxScaler()* class of the *sklearn* library as given in equation (1). This technique is used as our dataset is not Gaussian and we need to preserve the distribution of data.

$$X_{new} = (X - X_{min}) / (X_{max} - X_{min}), \quad (1)$$

where:
X = some dataset value we are about to rescale;
X_{new} = rescaled value corresponds to the X;
X_{min} = minimum value in the dataset;
X_{max} = maximum value in the dataset.

Model training and testing

Here comes the major part, i.e., training and testing the model. What we do, we split the entire dataset into two parts –

- Training Set is the portion fed to the model for learning reasons.
- Testing Set is the portion used to test against what the model has learned from the training set.

The predicted values of the testing set are compared against the accurate values of the dataset to determine the models' accuracy. This data partitioning technique is needed to avoid overfitting. Overfitting is the situation when the model couldn't fit the dataset reliably. We use the *train_test_split()* class of the *scikit-learn* library for data splitting. Data splitting is done in 4:1, where 20% of the data are utilized for evaluation and testing, and 80% of the data are employed for training. In this paper, the performance of three models is compared– the RF Regression model, the ET Regressor model, and the ANN.

In the Model Training and Testing phase, we trained individual models for each crop type to ensure that predictions for one crop (e.g., maize) were not influenced by data from other crops (e.g., potatoes). The dataset was preprocessed to separate crop-specific data, allowing each model to focus exclusively on its respective crop, reducing cross-crop variability and enhancing prediction accuracy. Additionally, we performed an 80/20 train-test split for each crop model, ensuring that test data remained independent from the training set, with no data leakage. Future studies could benefit from investigating country-specific factors by training models separately on region-based subsets or incorporating country data as a feature to further refine crop yield predictions.

In this paper, we implement a *Feedforward ANN* using Root Mean Squared Propagation (*RMSprop)* as an *organizer* from TensorFlow. The model is trained with a *learning rate* of 0.001 for 50 epochs, and the *dataset is split* 4:1 for training and testing. Three hidden layers with 256, 128, and 64 nodes use the Rectified Linear Unit *(ReLU) activation function*. The output layer has a single node with a *'linear' activation function*.

Results and Discussion

The success of any model is tested on the basis of following performance evaluation metrics, such as, Mean Square Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). Its predictive accuracy is determined using various evaluation metrics, tied to deep learning tasks.

- Mean Square Error (MSE): Using the MSE, how well a regression line fits a set of data points can be determined as given in equation (2). It is a risk function representing the squared error loss expected value (Scikit Learn, a n.d.).

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - Y_i)^2 , \quad (2)$$

where y_i = observed values and Y_i = predicted values.

- Mean Absolute Error (MAE): MAE is the difference between the prediction and the true value of observation (Scikit Learn, b n.d.) as given in equation (3).

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - Y_i|, \quad (3)$$

- R-squared (R^2): The R-squared or coefficient of determination, indicates how much of the variance is explained by the independent variables (features or predictors) (Scikit Learn, c n.d.) as given in equation (4).

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - Y_i)^2}{\sum_{i=1}^{N} \left\{y_i - \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)\right\}^2}, \quad (4)$$

RF Regression implements *Aggregation and Bagging* and is considered the best when accuracy, efficiency, and performance are the main concern. The model is effective for implementation on large databases but with high computational costs. Moreover, the model provides *no interpretability* and frequent occurrences of *overfitting* problems (Chaya, 2020). In this paper, we use a *scatter plot* to examine the performance of the model as shown in Figure 5. How close the bubbles are to the line determines how accurate the model is, that is why we used scatter plots. The MSE, MAE and R^2 values of this method are tabulated in Table 1.

The ET Regression creates randomized decision trees, unlike RF Regression, and produces combined (average) output from the decision trees. The ET Regression Model gives the predictive accuracy and performance as of the RF Regression Model. Also, ET Regressor is a good and effective choice when the *computational cost* is a concern as well as *careful analysis* and *selection of features* are required. ET Regressor provides *low variance* (Budu, 2022; Thankachan n.d.). In this paper, *Scatter Plot is used* to examine the accuracy of the model at computational costs. It helps to display a set of data points measured for two different variables and to see if there are patterns in the data set as shown in Figure 6. X-axis represents the "Exp Logs" values that showcases values of the experimental data points on the horizontal axis while the Y-axis represents predicted or calculated values corresponding to the experimental data points on the vertical axis. Exp Logs likely stands for "Experimental Logarithmic Values." This suggests that the data points on the x-axis have been transformed using a logarithmic function. The MSE, MAE, and R^2 values of this method are tabulated in Table 1.

The general form of a linear regression model is given in equation 5:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon, \quad (5)$$

where:
y is the dependent variable;
$x_1, x_2, ..., x_p$ are the independent variables;
β_0 is the intercept (the value of y when all x variables are 0);
$\beta_1, \beta_2, ..., \beta_p$ are the coefficients of the independent variables, which measure the change in y for a one-unit change in the corresponding x;
ε is the error term, representing the variation in y that cannot be explained by the x variables.

The form of the relationship between y and the x variables is typically linear, but could be polynomial or another form in more complex models.

ANNs are human-brain-inspired and parallel processing models in which information is stored in a distributive fashion (in nodes) all over the network. ANN might prove effective when there is incomplete information, non-linear and complex relationships, and highly volatile information (Artificial Neural Network Tutorial n.d.). The scatter plot for ANN is shown in Figure 7. The MSE, MAE and R^2 values of this method are tabulated in Table 1. In this paper, *Line Plots* and *Scatter Plot* have been used to represent the *loss* and *accuracy* of the ANN model.

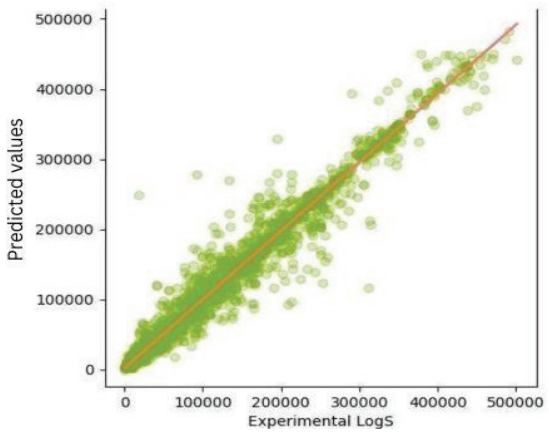


Figure 5. Scatter plot of RF regression.

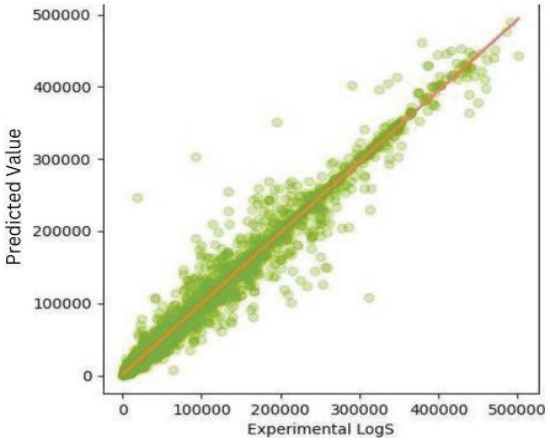


Figure 6. Scatter plot of ET regression.

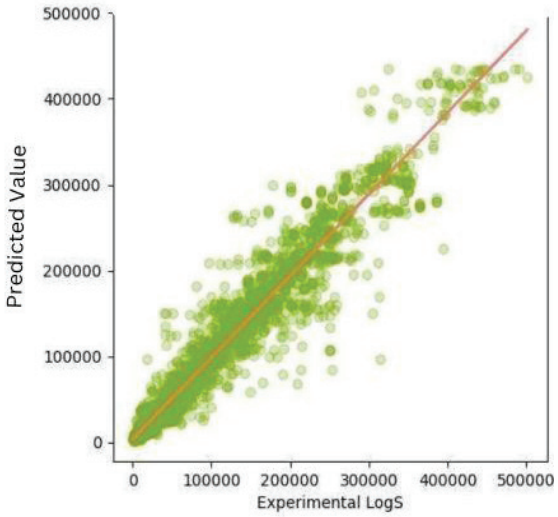
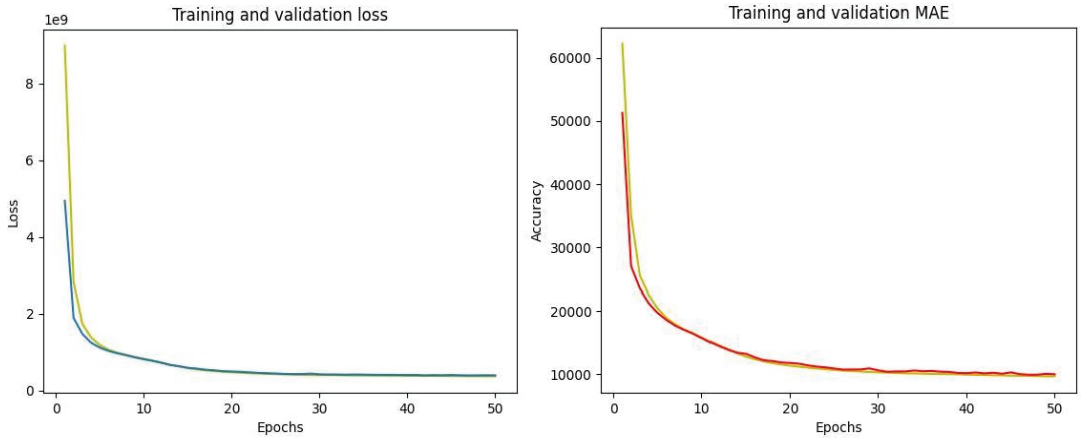


Figure 7. Scatter plot of ANN.

Figure 8 (a) illustrates a typical training process where the model's loss decreases with each epoch, indicating improved performance. The corresponding accuracy plot in Figure 8 (b) shows a steady increase in accuracy as the model learns from the data. The model seems to be performing well up until approximately epoch 40-50, after reaching its highest accuracy point, the model's performance began to degrade, possibly due to overfitting or other factors. Hence, both graphs show positive results.

Furthermore, hyperparameter tuning was performed using Keras Tuner to optimize the configuration of the neural network model. For hyperparameter tuning, we utilized a Random Search approach to optimize the architecture and training parameters of our neural network model for crop yield prediction. The hyperparameters tuned include the number of layers in the model, which varied between 2 and 50 layers, and the number of units in each dense layer, with values ranging from 32 to 256 in steps of 32. Additionally, we experimented with three different learning rates: 0.01, 0.001, and 0.0001. The tuning process was configured with a maximum of 5 trials, each trial being executed 3 times to ensure robustness. It is important to note that the results obtained from the hyperparameter tuning process are not subjected to traditional statistical analysis. However, the hyperparameter optimization process integrates techniques such as cross-validation and random search to systematically explore the hyperparameter space. The objective of this optimization process is to identify the optimal configuration that minimizes the mean absolute error on the validation dataset, thereby enhancing the generalization performance of the model. While formal statistical tests are not directly applied to the results, the utilization of these



(a) (b)

Figure 8. Line plots of (a) loss and (b) accuracy of ANN.

Table 1

Comparative performance for several evaluation metrics using various regression models.

Models	MSE	MAE	R^2
RF Regression	187388680.58	5662.69	0.9744
ET Regression	183648586.88	5249.03	0.9750
Artificial Neural Network	373094464.00	9811.87	0.9873

systematic techniques aims to improve the reliability and effectiveness of the model.

Analyzing the order of MSE and MAE values for all the regression models in Table 1, the ET Regression model has the least values, followed by RF Regression with slight increments, and then ANN with quite higher values. ANN has the highest R^2 values among all the algorithms i.e., 0.9873, but almost double the values of MAE and MSE as compared to that of the RF and ET Regression models. Apart from having the maximum R^2 value, it does not yield the best results because it shows the maximum errors with the values 373094464.00 of MSE and 9811.87 of MAE. The two high values of error overpower the highest accuracy of ANN. For appropriate results, MAE must be the least. Hence, we can say that the best regression model for crop yield prediction is the ET Regression model with a minimum MAE of 5249.03 and MSE of 183648586.88. However, our results are discussed in view of literature data to contextualize and interpret our findings within the broader research landscape.

Conclusion

In the context of modern agricultural challenges, accurate crop prediction is vital for enhancing food security, maximizing resource efficiency, and supporting sustainable farming practices. This paper compared three machine-learning models—RF Regression, ET Regressor, and Artificial Neural Network—for crop yield prediction. The ET Regression model demonstrated superior performance, achieving the lowest MAE (5249.03) and MSE (183648586.88), coupled with a high R^2 score, resulting in 97.5% accuracy. Despite the ANN model attaining the highest R^2 value, its higher MSE and MAE rendered it less suitable. These findings underscore the critical need to balance computational costs and model performance when selecting regression models. Consequently, the ET Regression model emerges as the most effective option for precise crop yield prediction. Future research should investigate ensemble techniques or hybrid models to enhance accuracy further, as well as incorporate remote sensing data and weather forecasts to refine predictions.

References

n.d. *Artificial Neural Network Tutorial.* javaTpoint. https://www.javatpoint.com/artificial-neural-network.

Batool, D., Shahbaz, M., Shahzad Asif, H., Shaukat, K., Alam, T.M., Hameed, I.A., ... & Luo, S. (2022). A hybrid approach to tea crop yield prediction using simulation models and machine learning. *Plants*, *11*(15), 1925. DOI: 10.3390/plants11151925

Bodapati, N., Himavaishnavi, J., Rohitha, V., Jagadeeswari, D.L., & Bhavana, P. (2022). Analyzing crop yield using machine learning. In *2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India,* (pp. 1-8). IEEE. DOI: 10.1109/ICEARS53579.2022.9752242.

Bonthu, H. (2021). *Analytics Vidhya.* 21 May. https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers- treating-the-odd-one-out/.

Budu, E. (2022). *Random Forest Vs. Extremely Randomized Trees.* Baeldung. 1 August. https://www.baeldung.com/cs/random-forest-vs-extremely- randomized-trees.

Chandraprabha, M., & Dhanaraj, R.K. (2021). Soil based prediction for crop yield using predictive analytics. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India,* (pp. 265-270). IEEE. DOI: 10.1109/ICAC3N53548.2021.9725758.

Chaya (2020). *Random Forest Regression.* Level Up Coding. 9 June. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84.

Dahikar, S.S., & Rode, S.V. (2014). Agricultural crop yield prediction using artificial neural network approach. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, *2*(1), 683-686. DOI: 7c68a32212c1f86f535f4c1658ff68399d0a9ddd.

Dharwadkar, N.V., Kalmani, V.H., & Thapa, V. (2023). Crop yield prediction using deep learning algorithm based on CNN-LSTM with attention

layer and skip connection. *Preprint*. DOI: 10.21203/rs.3.rs-3118781/v1.

Elavarasan, D., & Vincent, P.D.R. (2021). A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10009-10022. DOI: 10.1007/s12652-020-02752-y.

Feng, P., Wang, B., Li Liu, D., Waters, C., Xiao, D., Shi, L., & Yu, Q. (2020). Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. *Agricultural and Forest Meteorology*, *285*, 107922. DOI: 10.1016/j.agrformet.2020.107922.

Fritz, S., See, L., Bayas, J. C. L., Waldner, F., Jacques, D. C., Becker-Reshef, I., ... & McCallum, I. (2019). A comparison of global agricultural monitoring systems and current gaps. Agricultural Systems, 168, 258-272. DOI: 10.1016/j.agsy.2018.05.010.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*, 3-42. DOI: 10.1007/s10994-006-6226-1.

Ghai, D., Tripathi, S. L., Saxena, S., Chanda, M., & Alazab, M. (Eds.). (2022). *Machine Learning Algorithms for Signal and Image Processing*. John Wiley & Sons. DOI: 10.1002/9781119861850.

Gopal, P.M., & Bhargavi, R. (2019). A novel approach for efficient crop yield prediction. *Computers and Electronics in Agriculture*, *165*, 104968. DOI: 10.1016/j.compag.2019.104968.

Huilgol, P. (2020). Feature transformation and scaling techniques to boost your model performance. https://www.analyticsvidhya.com/blog/2020/07/types-of-feature-transformation- and-scaling/.

Iniyan, S., Varma, V.A., & Naidu, C.T. (2023). Crop yield prediction using machine learning techniques. *Advances in Engineering Software*, *175*, 103326. DOI: 10.1016/j.advengsoft.2022.103326.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., ... & Kim, S.H. (2016). Random forests for global and regional crop yield predictions. *PloS one*, *11*(6), e0156571. DOI: 10.1371/journal.pone.0156571.

Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop yield prediction using machine learning and deep learning techniques. *Procedia Computer Science*, *218*, 406-417. DOI: 10.1016/j.procs.2023.01.023.

Kanyutu, K. (2023). *Exploratory Data Analysis Ultimate Guide.* DEV Community. 26 February. https://dev.to/kim_kanyutu/exploratory-data-analysis- ultimate-guide-35e1.

Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, *10*, 452963. DOI: 10.3389/fpls.2019.00621.

Khaki, S., Wang, L., & Archontoulis, S.V. (2020). A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, *10*, 492736. DOI: 10.3389/fpls.2019.01750.

Kim, N., Ha, K.J., Park, N.W., Cho, J., Hong, S., & Lee, Y.W. (2019). A comparison between major artificial intelligence models for crop yield prediction: Case study of the midwestern United States, 2006–2015. *ISPRS International Journal of Geo-Information*, *8*(5), 240. DOI: 10.3390/ijgi8050240.

Kumar, Y.J.N., Spandana, V., Vaishnavi, V.S., Neha, K., & Devi, V.G. R. R. (2020). Supervised machine learning approach for crop yield prediction in agriculture sector. In *2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India,* (pp. 736-741). IEEE. DOI: 10.1109/ICCES48766.2020.9137868.

Kuwata, K., & Shibasaki, R. (2016). Estimating corn yield in the United States with MODIS EVI and machine learning methods. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *3*, 131-136. DOI: 10.5194/isprsannals-III-8-131-2016.

Mishra, S., Mishra, D., & Santra, G. H. (2016). Applications of machine learning techniques in agricultural crop production: a review paper. *Indian Journal of Science and Technology*, 9(38), 1-14. DOI: 10.17485/ijst/2016/v9i38/95032.

Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019). Crop yield prediction using machine learning algorithms. In *2019 5th International Conference on Image Information Processing (ICIIP),* Shimla, India, (pp. 125-130). IEEE. DOI: 10.1109/ICIIP47207.2019.8985951.

Nithya, V., Josephine, M.S., & Jeyabalaraja, V. (2023). IoT-based crop yield prediction system in Indian sub-continent using machine learning techniques. *Remote Sensing in Earth Systems Sciences*, *6*(3), 156-166. DOI: 10.1007/s41976-023-00097-6.

Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Hybrid deep learning-based models for crop yield prediction. *Applied Artificial Intelligence*, *36*(1), 2031822. DOI: 10.1080/08839514.2022.2031823.

Pandith, V., Kour, H., Singh, S., Manhas, J., & Sharma, V. (2020). Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. *Journal of Scientific Research*, *64*(2), 394-398. DOI: 10.37398/JSR.2020.640254.

Patel, K., & Patel, H.B. (2021). A comparative analysis of supervised machine learning algorithm for agriculture crop prediction. In

2021 4th International Conference on Electrical, Computer and Communication Technologies (ICECCT), Erode, India, (pp. 1-5). IEEE. DOI: 10.1109/ICECCT52121.2021.9616731.

Patel, R. n.d. *Crop Yield Prediction Dataset.* Kaggle. https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset.

Satir, O., & Berberoglu, S. (2016). Crop yield prediction under soil salinity using satellite derived vegetation indices. *Field Crops Research*, *192*, 134-143. DOI: 10.1016/j.fcr.2016.04.028.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117. DOI: 10.1016/j.neunet.2014.09.003.

n.d. *Scikit learn.* https://scikitlearn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html#sklearn.metrics.mean_absolute_error.

n.d. *Scikit Learn.* https://scikit- learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.

n.d. *Scikit Learn.* https://scikit- learn.org/stable/modules/generated/sklearn.metrics.r2_score.html#sklearn.metric s.r2_score.

Segal, M.R. (2004). Machine Learning Benchmarks and Random Forest Regression. UCSF: Center for Bioinformatics and Molecular Biostatistics. Retrieved from https://escholarship.org/uc/item/35x3v9t4.

Seldon. (2021). *Machine Learning Regression.* 29 October. https://www.seldon.io/machine-learning-regression-explained#:~:text=Regression%20is%20a%20technique%20for,used%20to%20predict%20continuous%20outcomes.

Singh, G. (2021). *Introduction to Artificial Neural Networks.* Analytics Vidhya. 6 September. https://www.analyticsvidhya.com/blog/2021/09/introduction-to- artificial-neural-networks/.

Sreerama, A.S., & Sagar, B.M. (2020). A machine learning approach to crop yield prediction. *International Research Journal of Engineering and Technology (IRJET)* 07 (05): 4, 6616-6619.

Suresh, N., Ramesh, N.V.K., Inthiyaz, S., Priya, P.P., Nagasowmika, K., Kumar, K.V.H., ... & Reddy, B.N. K. (2021). Crop yield prediction using random forest algorithm. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, *Coimbatore, India,* (pp. 279-282). IEEE. DOI: 10.1109/ICACCS51430.2021.9441871.

Swain, D., Lakum, S., Patel, S., & Patro, P. (2024). An Efficient Crop Yield Prediction System Using Machine Learning. *EAI Endorsed Transactions on Internet of Things*, 10, 1-5. DOI: 10.4108/eetiot.5333.

Thankachan, K. n.d. (2022). *What? When? How? ExtraTrees Classifier.* Towards Data Science. https://towardsdatascience.com/what-when-how-extratrees-classifier- c939f905851c.

Venugopal, A., Aparna, S., Mani, J., Mathew, R., & Williams, V. (2021). Crop yield prediction using machine learning algorithms. *International Journal of Engineering Research & Technology (IJERT)*, 9(13), 87-91. DOI: 10.17577/IJERTCONV9IS13019.