Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enhancing Agricultural Forecasting with an Ensemble Learning Approach for Broccoli Yield Prediction

[1, *] G.Alexandar Narkunam, [2]K.Kala, [3] S.ArunPandiyan
[1,3] *Department of Computer Science, Alagappa University, Karaikudi, India*
[2]*Department of Computer Science, Nachiappa Swamigal Arts and Science College, Koviloor, India*
Email: [1]narkunam2015@gmail.com[*], [2]kasinathan71@gmail.com, [3] spandiyan01@gmail.com
[*] *Corresponding author*

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Accurate yield prediction for broccoli remains a critical challenge due to the complex, non-linear interactions between climatic variables, soil properties, and agronomic practices. Traditional statistical models often fail to capture these interactions, leading to suboptimal decision-making and resource inefficiencies for farmers. To overcome these limitations, this study proposes an advanced machine learning-based approach called the Broccoli Yield Prediction Ensemble Method (BYPEM), designed to improve prediction accuracy and agricultural planning. BYPEM integrates both bagging and boosting ensemble learning techniques for robust broccoli yield forecasting. The study begins with a comprehensive data preprocessing phase, including handling missing values, outlier removal, categorical variable encoding, and normalization. Feature selection is performed using backward elimination to retain the most relevant predictors. The dataset is split into training and test sets through stratified sampling to ensure balanced representation. In the model development phase, BYPEM applies bagging methods such as Random Forest Regressor and Extra Trees Regressor to reduce variance, and boosting methods such as Gradient Boosting Regressor (GBR), XGBoost, LightGBM, and CatBoost to minimize bias by iteratively improving predictions. Hyperparameter tuning further optimizes model performance. The models are evaluated using multiple metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 score, and overall accuracy. Among all models, XGBoost achieves the highest predictive performance, confirming the effectiveness of the BYPEM framework in capturing complex yield dynamics. This research demonstrates how ensemble learning can support sustainable agriculture by enhancing precision in broccoli yield prediction and guiding data-driven farm management strategies.

Keywords: Broccoli yield prediction, machine learning models, ensemble learning techniques, BYPEM, feature selection, bagging and boosting methods, hyperparameter tuning.

1. INTRODUCTION

Agricultural productivity, particularly crop yield prediction, is crucial for ensuring food security, optimizing resources, and improving farming practices [1, 2]. Accurate yield prediction helps make informed decisions on irrigation, fertilization, pest control, and harvesting. This research develops a machine learning model to predict broccoli yield (kg) based on environmental, soil, and agricultural factors such as soil moisture, temperature, rainfall, and market price. By using data science techniques, the study aims to provide accurate predictions to help farmers improve efficiency. Predicting crop yield, especially for broccoli, involves datasets with both numerical and categorical features. The study applies preprocessing steps like handling missing values, removing outliers, and normalizing data, followed by feature selection. Models like Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM, and CatBoost are trained and tested to assess their accuracy. The study focuses on key factors such as pH, moisture content, nitrogen levels, temperature, and rainfall, comparing model performance based on predictive accuracy and other metrics. A significant challenge in yield prediction is the variability of factors like soil conditions, climate, and farming practices. For broccoli, environmental conditions like temperature, rainfall, and

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

moisture content are crucial, along with soil quality factors like nitrogen, phosphorus, and potassium levels. Traditional methods often use simpler models, which may not capture the complex interactions between these variables [3]. Studies have shown that limiting lateral branches and managing pests like Diamondback Moth (DBM) can affect yield [4, 5]. Karso et al. highlighted that broccoli (*Brassica oleracea* L.), though suitable for year-round cultivation under greenhouse conditions, is vulnerable to several insect pests including Brevicoryne brassicae, Pieris rapae, Plutella xylostella, and Agrotis ipsilon [20]. The study demonstrated that the use of sticky traps effectively reduced pest infestation levels, with infected percentages significantly lowered (to 0–4%) across pest types. This method improved both yield quantity (up to 1200 kg/greenhouse) and quality, while eliminating the need for chemical pesticides, making it a sustainable solution within integrated pest management strategies. This research aims to bridge this gap by applying machine learning to capture these complex relationships. The study includes data preprocessing, feature selection, and model evaluation. Models are trained on a dataset split into training, validation, and test sets. Recent studies on broccoli production optimization under climate change [6, 7] highlight the importance of improving yield predictions through machine learning. Machine learning techniques, especially regression models, have shown promise in overcoming the limitations of traditional methods. For instance, studies have found that XGBoost outperforms other models in terms of accuracy and efficiency [8]. Random Forest and Extra Trees Regressor models are also popular for reducing overfitting and providing valuable feature importance metrics. The methodology involves handling missing values, removing outliers, normalizing numerical features, and encoding categorical variables.

Backward elimination removes non-significant features, ensuring model reliability in predicting broccoli . Ciancaleoni et al. conducted a reduced rank factorial regression analysis to evaluate the impact of environmental variables—particularly nitrogen content, rainfall, minimum temperature, and clay content—on broccoli yield under Organic Agriculture (OA) and Low Input (LI) systems. These factors explained 91% of the G×E interaction, highlighting the importance of managing key environmental conditions and selecting genotypes adapted to specific pedo-climatic contexts in sustainable agriculture [14]. Kim et al. used the ALMANAC model to simulate the impact of climate change and cropping practices on broccoli yield in California. Through 33,600 simulations combining climate scenarios, CO_2 levels, nitrogen rates, and plant densities, they found nitrogen application had the greatest influence on yield, followed by CO_2 levels. Under stress conditions, low density and low nitrogen inputs maximized yield, while elevated CO_2 and temperature improved responses to higher inputs. The study highlights the need for adaptive strategies to sustain broccoli production under changing climate conditions. Johansen et al. investigated how latitude-related climatic factors affect the sensory quality of broccoli florets across Europe. Broccoli grown under low temperatures and long daylight hours in northern regions showed enhanced attributes like bud coarseness, uniform color, crispness, and juiciness, while southern sites with higher temperatures and shorter days exhibited increased bitterness and stale flavor. Their findings highlight the influence of temperature and light interactions on sensory traits and support region-specific marketing of broccoli based on climatic growing conditions [16]. Sola et al. examined how low (LT) and high (HT) growing temperatures affect the phytochemical profile and bioactivity of broccoli microgreens. LT conditions enhanced total phenolics, tannins, glucosinolates, and sinapic acid, indicating improved nutritional value. In contrast, HT elevated soluble sugars and indole-3-acetic acid, suggesting stronger osmotic stress and defense responses. Both temperature stresses reduced chlorophyll and antioxidant potential. While both LT and HT increased α-amylase inhibition, only LT improved lipase inhibition. The study concludes that LT-grown broccoli microgreens are more nutritionally beneficial than those grown under HT conditions [17]. Scuderi et al. conducted an environmental and economic assessment of an innovative organic broccoli cultivation model proposed by the Bresov project in Sicily. Using Life Cycle Assessment and gross income evaluation, the study revealed a 60–100% reduction in environmental impact compared to conventional methods. Although organic yields were slightly lower and production costs higher—leading to a 61% reduction in gross income—these losses were offset by Common Agricultural Policy (CAP) subsidies and premium market prices. The Bresov protocol, incorporating organic fertilizers, natural crop protection, and improved soil management, demonstrated potential for enhancing sustainability, environmental protection, and farmer profitability in organic broccoli farming [18]. Wang et al. emphasized the need for innovation in broccoli cultivation to meet increasing market demands for quality and yield [19]. The study proposed an integrated approach involving advancements in seedling technology, transplanting methods, nutrient and pest management, and the adoption of precision

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

agriculture and green circular production models. Results demonstrated that these integrated innovations significantly enhance yield, quality, and production efficiency. The research highlights the potential of smart agriculture and sustainable practices to support the broader development of the broccoli industry, advocating for future improvements in digital management and ecological farming systems. This study contributes to agricultural science by applying machine learning for yield prediction. By using advanced bagging and boosting techniques, the research aims to provide more accurate predictions than traditional methods. The preprocessing and feature selection techniques used can be applied to other crop yield prediction tasks, benefiting various agricultural sectors. Research on crop and water productivity for broccoli under irrigation systems highlights the importance of water management strategies in optimizing yield [9]. The study emphasizes hyperparameter tuning, model selection, and ensemble methods to achieve high prediction accuracy, helping farmers make better decisions and promote sustainable farming practices. XGBoost is expected to outperform other models, providing a reliable tool for forecasting broccoli yield. The study seeks to identify key factors influencing yield, such as soil moisture, temperature, and market price, and improve yield prediction. Enhanced predictions can help farmers optimize practices, increase productivity, and contribute to food security, while advancing precision agriculture for more sustainable farming practices.

2. BYPEM PROPOSED MODEL

2.1 Dataset

The dataset used in this study has been sourced from ICAR (Indian Council of Agricultural Research) and has represented agricultural data from the Tamil Nadu region in India, as shown in Figure 1. It has consisted of 473 records with key attributes essential for broccoli yield analysis. These attributes have included Soil_Type, pH, Moisture_Content, Nitrogen_Content, Phosphorus_Content, Potassium_Content, Water_Irrigation_L, Fertilizer Inputs (Nitrogen, Phosphorus, Potassium in kg), Temperature (Spring, Summer, Fall), Rainfall (Spring, Summer, Fall in mm), Broccoli_Yield_kg, and Market_Price_INR**.** The dataset has provided valuable insights into soil properties, climatic conditions, and fertilizer usage, making it a useful resource for machine learning applications in agricultural optimization and yield prediction**.**

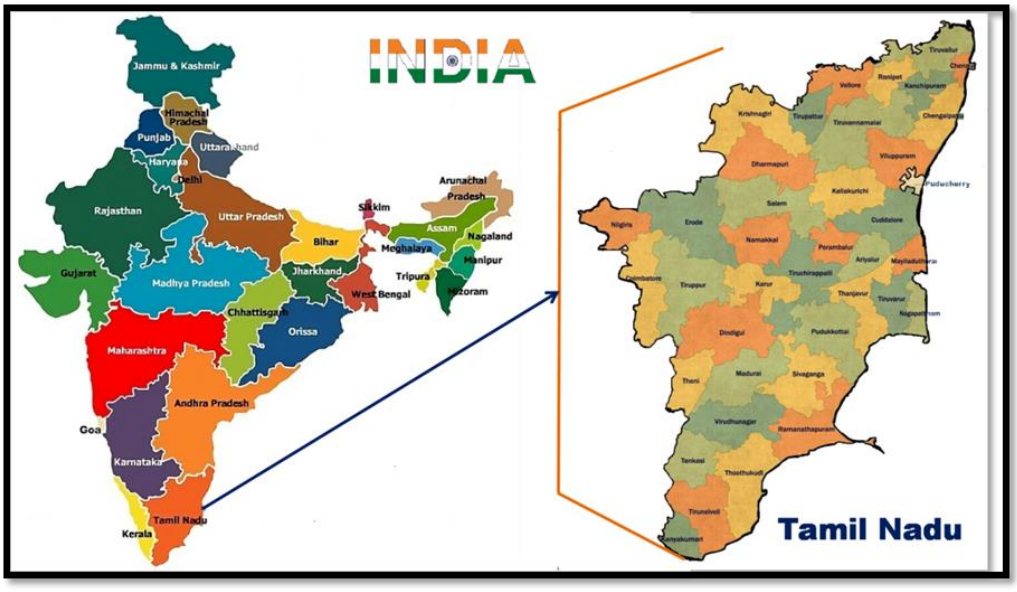

Figure 1. Dataset Representation for Tamil Nadu Region

Data preprocessing is the initial step before building a machine learning model. The first task is to handle missing values in the dataset. For numerical columns such as pH, Moisture_Content, Nitrogen_Content, Phosphorus_Content, Potassium_Content, Temperature_Spring_C, Temperature_Summer_C, Temperature_Fall_C, Rainfall_Spring_mm, Rainfall_Summer_mm, Rainfall_Fall_mm, and Market_Price_INR, KNN imputation has been used to fill the missing values. If KNN imputation is not suitable, the median value has

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

been used instead. The median is calculated differently depending on whether the number of data points is odd or even. The overall workflow of the proposed model is illustrated in Figure 2. Equation 1 represents the calculation of the median for datasets with odd and even numbers of data points.

$$Median = X_{\frac{n+1}{2}} \text{ and } Median = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} \quad (1)$$

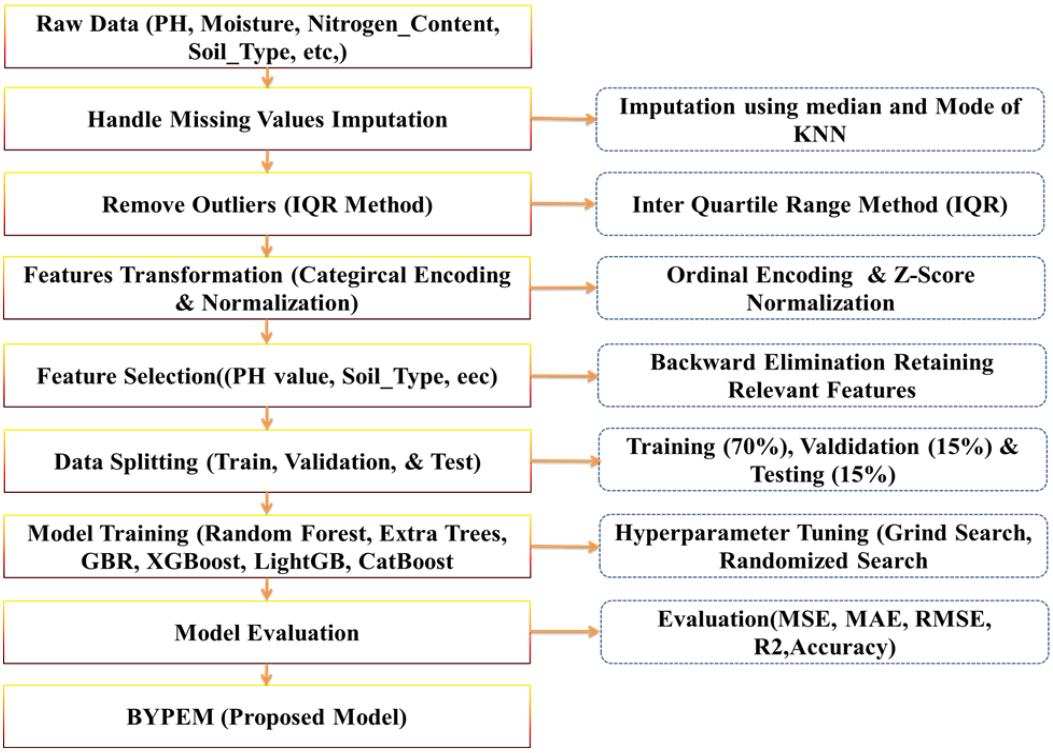

Figure 2. Architecture of Proposed Ensemble Model

For categorical columns such as Soil_Type, missing values have been imputed using the mode (the most frequent value), which is consistent with the most common type of soil in the dataset, ensuring that missing data does not disrupt the consistency. The next preprocessing step is handling outliers. The Inter quartile Range (IQR) method has been used to detect and remove outliers from numerical features. The IQR is calculated based on equation 2 as,

$$IQR = Q3 - Q1 \quad (2)$$

Where Q1 and Q3 are the first and third quartiles, respectively. Outliers are identified using the following bounds,

$$Lower\ Bound = Q1 - 1.5\ X\ IQR \text{ and Upper Bound = Q3+1.5 X IRQ} \quad (3)$$

Data points falling outside these bounds have been removed from the dataset. For categorical variables, Soil_Type has been transformed into numerical values using Ordinal Encoding, assigning integer values to each category (e.g., Sandy = 1, Loamy = 2, Clay = 3). Finally, Z-score normalization has been applied to numerical features such as pH, Moisture_Content, Nitrogen_Content, Phosphorus_Content, Potassium_Content, Broccoli_Yield_kg, Water_Irrigation_L, Nitrogen_Fertilizer_kg, Phosphorus_Fertilizer_kg, Potassium_Fertilizer_kg, Temperature_Spring_C, Temperature_Summer_C, Temperature_Fall_C, Rainfall_Spring_mm, Rainfall_Summer_mm, Rainfall_Fall_mm, and Market_Price_INR. This normalization transforms each feature to have a mean of 0 and a standard deviation of 1,

$$Z = \frac{X - \mu}{\sigma} \quad (4)$$

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Where X is the data point, μ is the mean, and σ is the standard deviation of the feature.

2.2 Feature Selection

After preprocessing, feature selection is performed to identify the most relevant features for predicting Broccoli_Yield_kg. This step involves using backward elimination, a wrapper method. Initially, a linear regression model has been trained using all features, and the model's performance has been evaluated using the R-squared equation metric 5,

$$R^2 = 1 - \frac{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2} \quad (5)$$

Where y_i is the actual value of Broccoli_Yield_kg, $\hat{y}_i$ is the predicted value from the model, $\bar{y}$ is the mean of the actual values. The p-value for each feature has been calculated using a statistical test (usually the t-test in linear regression). Features such as Rainfall_Fall_mm have been identified as least significant, and features with p-values greater than a chosen threshold (0.05) have been removed. This process has been repeated iteratively until no more features could be eliminated without adversely affecting the model performance. The features that have been retained for model training include pH, Moisture_Content, Nitrogen_Content, Potassium_Content, Temperature_Spring_C, Temperature_Summer_C, Market_Price_INR, and the encoded Soil_Type.

2.3 Data Splitting

Once the relevant features have been selected, the dataset has been divided into training **and** test sets using stratified sampling. The target variable Broccoli_Yield_kg has been preserved proportionally across all splits. The dataset has been divided as training set 70% of the data, **t**esting set 30% of the data.

This splitting process has been carried out using Python's train_test_split function, with a fixed random seed (random_state=42) to ensure reproducibility. The training set is used to train the machine learning models, while the validation set is used to fine-tune hyperparameters and avoid overfitting. The test set is reserved for the final model evaluation to assess its generalization performance on unseen data.

2.4 Bagging and Boosting Techniques

The next phase of the model development involves applying machine learning models based on bagging and boosting techniques, focusing on Random Forest Regressor and Extra Trees Regressor for bagging, and Gradient Boosting Regressor (GBR**)**, XGBoost, LightGBM, and CatBoost for boosting. These models are designed to reduce variance and improve predictive accuracy by combining multiple models' outputs.Bagging-based Models like Random Forest Regressor (RF) build multiple decision trees and average their predictions to reduce variance. The decision trees in Random Forest are trained on bootstrapped samples, meaning each tree sees a slightly different view of the data. In the case of Random Forest, key attributes such as pH, Moisture_Content, Nitrogen_Content, Potassium_Content, Temperature_Spring_C, Temperature_Summer_C, Market_Price_INR, and Soil_Type (encoded) are used to train the model. The prediction from the Random Forest model is calculated by averaging the predictions of all trees, using the equation 6:

$$\hat{y}\,\mathrm{RF} = \frac{1}{n}\sum_{i-1}^{n}\hat{y}_i \quad (6)$$

where n is the number of trees, and $\hat{y}_i$ is the prediction from the i-th tree. Hyperparameter tuning for Random Forest involves adjusting parameters such as n_estimators, max_depth, min_samples_split, and min_samples_leaf to optimize the model's performance using techniques like Grid Search or Randomized Search. The Extra Trees Regressor (ET) model is similar to Random Forest, but with an added layer of randomness. Extra Trees selects random thresholds for splits, making training faster while potentially improving model generalization. Key attributes such as Temperature_Spring_C, Moisture_Content, Nitrogen_Content, Market_Price_INR, and Soil_Type (encoded) are also used in Extra Trees. The model builds multiple trees, and the prediction is made by

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

averaging the outputs from all trees, much like Random Forest. Hyperparameter tuning for Extra Trees focuses on parameters such as n_estimators, max_depth, and min_samples_split.

In Boosting-based Models, the goal is to reduce bias by iteratively training models that correct the errors made by the previous models. The Gradient Boosting Regressor (GBR) builds trees sequentially, where each new tree aims to correct the residual errors from the previous trees. Attributes such as pH, Moisture_Content, Potassium_Content, Temperature_Spring_C, Temperature_Summer_C, and Market_Price_INR are essential for capturing the relationship between these features and Broccoli yield. The equation 7 for updating predictions in GBR is given by

$$f_m(x) = f_{m-1}(x) + \eta . h_m(x) \quad (7)$$

where $f_m(x)$ is the prediction at the m-th iteration, f_{m-1} is the prediction from the previous iteration, η is the learning rate, and $h_m(x)$is the new decision tree model. Hyperparameters like learning_rate, n_estimators, max_depth, and subsample are tuned to optimize model performance.

XGBoost is a highly optimized version of gradient boosting that is efficient in handling large datasets. It is known for its speed, performance, and ability to incorporate regularization, which helps prevent overfitting. Important features used in XGBoost include Rainfall_Spring_mm, Temperature_Spring_C, Market_Price_INR, and Potassium_Content. Hyperparameter tuning for XGBoost includes adjusting learning_rate, n_estimators, subsample, and colsample_bytree to fine-tune the model.

LightGBM is a scalable version of boosting known for its efficiency with large datasets. It uses histogram-based algorithms to optimize memory usage and reduce training time. Moisture_Content, Nitrogen_Fertilizer_kg, Temperature_Summer_C, and Market_Price_INR are the key attributes leveraged in LightGBM. Hyperparameters like num_leaves, learning_rate, and max_depth are adjusted to improve the model's performance. Finally, CatBoost is particularly effective in handling categorical variables such as Soil_Type, which can have a significant impact on crop yield. CatBoost automatically handles categorical variables without requiring extensive preprocessing, which simplifies the model-building process. Key features used in CatBoost include Soil_Type (encoded), Moisture_Content, Temperature_Spring_C, and Market_Price_INR. The model's hyperparameters, such as iterations, learning_rate, and depth, are tuned to optimize model performance. Each of these boosting models works iteratively, correcting errors from previous models, and focuses on minimizing bias by leveraging different attributes of the dataset. The hyperparameter tuning in both bagging and boosting models ensures that the best configuration of each model is achieved, improving the overall prediction accuracy for Broccoli yield. Together, the bagging and boosting techniques, utilizing important attributes such as pH, Moisture_Content, Nitrogen_Content, Potassium_Content, Temperature_Spring_C, Temperature_Summer_C, Market_Price_INR, and Soil_Type (encoded), create an ensemble of models that work together to reduce variance, improve generalization, and enhance the accuracy of Broccoli yield predictions. This hybrid approach combines the strengths of both techniques, providing a robust framework for predicting Broccoli_Yield_kg.

BYPEM ALGORITHM

Dataset D = {X, y}, where

X = {Soil_Type, PH, Moisture_Content, ...} (Features)

y = Broccoli_Yield (Target variable)

y_pred (Predicted Broccoli Yield)

Begin

Load dataset D containing features X and target variable y

Xnum = Impute(KNN or Median) for numerical features

Xcat = Impute(Mode) for categorical features

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

```
    Xnum = Remove_Outliers_IQR(Xnum)
    Xcat = Ordinal_Encode(Xcat)
    FOR each feature x in Xnum DO:
      x_norm = (x - mean(x)) / std(x)
    FOR each feature Xi in X DO:
    Pi = Compute_p_value(Xi)
    IF Pi > threshold THEN:
      Remove Xi
      X_selected = {Xi | Pi ≤ threshold}
      D_train = 70% of D
      D_test  = 30% of D
     Train Models(Random Forest , Extra Trees, Gradient Boosting, XGBoost, LightGBM, CatBoost )
  FOR each trained model DO:
    y_pred = Predict(D_test)
    R² = 1 - (Σ(y_true - y_pred)² / Σ(y_true - y_avg)²)
 Select the model with the highest R² score
 y_pred = Best_Model(D_test)
 Return y_pred
END
```

3. RESULT AND DISCUSSION

3.1 MAE (Mean Absolute Error)

The Mean Absolute Error (MAE) provides a measure of the average magnitude of errors between the real and predicted values, without considering their direction. The equation 8 for MAE is,

$$MAE = \frac{1}{n}\sum_{i-1}^{n}|y_i - \hat{y}_i| \quad (8)$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, n is the number of data points. A lower MAE indicates that the model's predictions are closer to the actual values. Based on the results, XGBoost performed the best with a MAE of 0.189, meaning its predictions were more accurate on average compared to models like Extra Trees Regressor (0.468) and Random Forest Regressor (0.452). This reflects the model's lower average error in prediction. Compared to other studies, XGBoost demonstrated stronger predictive performance with lower errors than many traditional models used in agricultural yield prediction tasks.

3.2 MSE (Mean Squared Error)

The Mean Squared Error (MSE) penalizes larger errors more severely than MAE, as it squares the differences between actual and predicted values. The equation 9 for MSE is:

$$MAE = \frac{1}{n}\sum_{i-1}^{n}(y_i - \hat{y}_i)^2 \quad (9)$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, n is the number of data points. A lower MSE indicates fewer large discrepancies between actual and predicted values. XGBoost showed the best performance with the lowest

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

MSE of 0.158, highlighting that its predictions contained fewer significant errors. This is in line with findings in recent studies where XGBoost consistently outperformed other algorithms in MSE when predicting agricultural yields. On the other hand, Extra Trees Regressor exhibited the highest MSE of 0.369, indicating that its predictions had a higher variance from the actual values.

3.3 RMSE (Root Mean Squared Error)

The Root Mean Squared Error (RMSE**)** is the square root of MSE, making it more interpretable as it is in the same units as the target variable. It is sensitive to large errors and reflects the average magnitude of error. The formula 10 for RMSE is,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

Where y_i is the actual value, $\hat{y}_i$is the predicted value, n is the number of data points. XGBoost emerged as the leader, achieving an RMSE of 0.238, which is lower than that of Random Forest Regressor (0.594) and Extra Trees Regressor (0.603). This suggests that XGBoost was able to make smaller, more consistent errors in its predictions, achieving more reliable overall performance. These results align with previous studies, where boosting methods like Gradient Boosting and XGBoost demonstrated significantly better RMSE values compared to tree-based models in agricultural yield prediction tasks.

3.4 R^2 (Coefficient of Determination)

The R^2 (Coefficient of Determination) measures how well the model's predictions fit the actual data, representing the proportion of variance in the target variable explained by the model. The equation 11 for R^2 is:

$$R^2 = 1 - \frac{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2} \tag{11}$$

Where $y_i$is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the actual values, n is the number of data points. The model's performance can be assessed using the R^2 value, which reflects how well the model explains the variance in the data. In this case, XGBoost demonstrated superior performance with an R^2 value of 0.91, explaining 91.5% of the variance in Broccoli Yield (kg). This is a significant improvement over other models, such as the Extra Trees Regressor, which had an R^2 value of 0.841, indicating a closer fit to the data. The higher R^2 value of XGBoost indicates its better ability to capture the underlying patterns in the data compared to the other models.

3.5 Accuracy

The Accuracy metric measures the percentage of predictions that match the actual values. This provides a good indication of the precision of the model in terms of how close its predictions are to the actual values. The equation 12 for accuracy is,

$$Accuracy = \frac{Number\ of\ correct\ Predictions}{Total\ Number\ of\ Predictions} = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

TP (True Positives) are correctly predicted positive instances. TN (True Negatives) are correctly predicted negative instances. FP (False Positives) are incorrectly predicted positive instances. FN (False Negatives) are incorrectly predicted negative instances. XGBoost achieved the highest accuracy at 92.4%, meaning most of its predictions were correct. This shows strong predictive precision. Extra Trees Regressor had a lower accuracy of 82%. XGBoost also performed better than models like CNN-DNN and CNN-RNN, which usually have accuracy percentages below 87%.

3.7 Training and Testing Phase

During the training phase, various machine learning models were tested using ensemble learning techniques, including bagging (Random Forest, Extra Trees) and boosting (XGBoost). The best hyperparameters were selected using Grid Search and Randomized Search, ensuring that each model was optimized for the dataset. The models

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

were evaluated based on standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 Score, and Accuracy. Table 1 presents the accuracy of different models in the training phase.

Table 1. Training Performance of XGBoost, Random Forest, and Extra Trees.

Metric	XGBoost (Training)	Random Forest (Training)	Extra Trees (Training)
MAE	**0.105**	0.423	0.442
MSE	**0.184**	0.321	0.351
RMSE	**0.231**	0.570	0.594
R^2 Score	**0.951**	0.872	0.856

The trained models have been evaluated on an unseen test set (30% of the total data) in the testing phase to measure their generalization performance. The results of this evaluation are provided in Table 2.

Table 2. Testing Performance of XGBoost, Random Forest, and Extra Trees.

Metric	XGBoost (Testing)	Random Forest (Testing)	Extra Trees (Testing)
MAE	**0.189**	0.452	0.468
MSE	**0.158**	0.346	0.369
RMSE	**0.238**	0.594	0.603
R^2 Score	**0.935**	0.841	0.836

Table 3. Performance Comparison of Different Models Based on Evaluation Metrics

Models	MSE	RMSE	MAE	R^2
CNN-DNN (RS-SFM)	0.70	0.28	0.18	0.87
CNN-XGBOOST (RS-SFM)	0.10	0.32	0.23	0.78
XGBOOST (RS)	0.10	0.33	0.24	0.78
CNN-RNN (RS-SFM)	0.11	0.34	0.26	0.78
XGBOOST (RS-SFM)	0.17	0.30	0.21	0.83
CNN-LSTM (RS-SFM)	0.18	0.43	0.30	0.67
BYPEM (Proposed Model)	0.10	0.26	0.17	0.91

Table 3 has presented a comparative analysis of various models based on the evaluation metrics: MSE, RMSE, MAE, and R^2. These metrics have been used to assess the predictive accuracy and reliability of each model. The BYPEM Model has demonstrated superior performance, achieving the lowest error values (MSE: 0.10, RMSE: 0.26, MAE: 0.17) and the highest R^2 (0.91), indicating its effectiveness in accurately capturing the underlying patterns in the data. Models integrated with RS-SFM preprocessing, such as CNN-XGBOOST (RS-SFM) and XGBOOST (RS-

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

SFM), have consistently outperformed their counterparts without RS-SFM. Conversely, CNN-LSTM (RS-SFM) has shown relatively poor performance with the lowest R^2 (0.67) and higher error values. This analysis has confirmed that the BYPEM Model is the most robust and reliable among the models evaluated for the given research task.

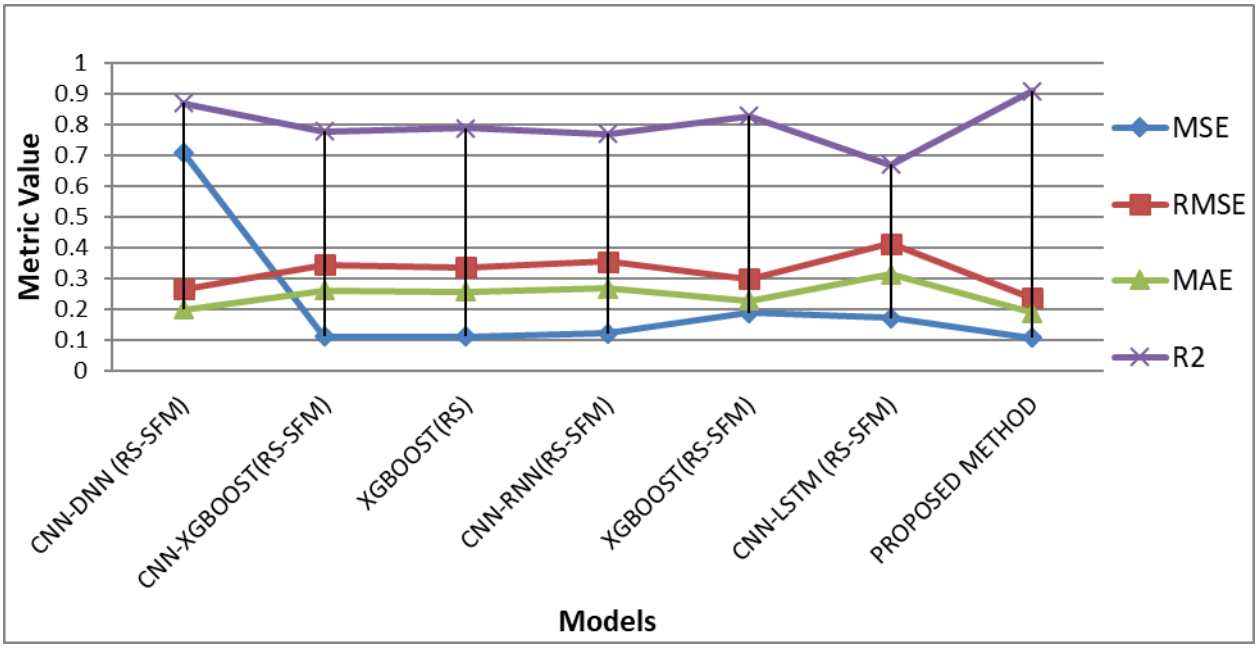

Figure 3: Performance Comparison of Machine Learning Models

The BYPEM Model performs better than other models by achieving lower error values and higher accuracy. As shown in Figure 3, it has the lowest MSE, RMSE, and MAE, indicating improved prediction accuracy, along with a higher R^2 score that demonstrates a stronger relationship between predicted and actual values. In terms of overall accuracy and performance, Figure 4 shows that the BYPEM Model surpasses all other approaches, making it the most effective model for this task.

Table 4: Accuracy Comparison of Different Models

Models	Accuracy (%)
CNN-DNN (RS–SFM)	88.3
CNN-XGBOOST (RS–SFM)	87.0
XGBOOST (RS)	90.3
CNN-RNN (RS–SFM)	87.8
XGBOOST (RS–SFM)	86.4
CNN-LSTM (RS–SFM)	89.0
BYPEM (Proposed Model)	**92.4**

The classification accuracy of various models has been evaluated to determine their effectiveness in handling the given task. As shown in Table 4, the BYPEM Model has achieved the highest accuracy of 92.4%, demonstrating its superior predictive capability over all other models. While XGBOOST (RS) has also performed well with 90.3% accuracy, other hybrid models such as CNN-LSTM (RS–SFM) and CNN-DNN (RS–SFM) have recorded accuracies of 89.0% and 88.3%, respectively. On the other hand, models like XGBOOST (RS–SFM) and CNN-XGBOOST (RS–

Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

SFM) have shown relatively lower performance, achieving 86.4% and 87.0% accuracy. These results, as summarized in Table 4, have highlighted the consistent and superior performance of the BYPEM Model in terms of classification accuracy, validating its suitability for the research objective.

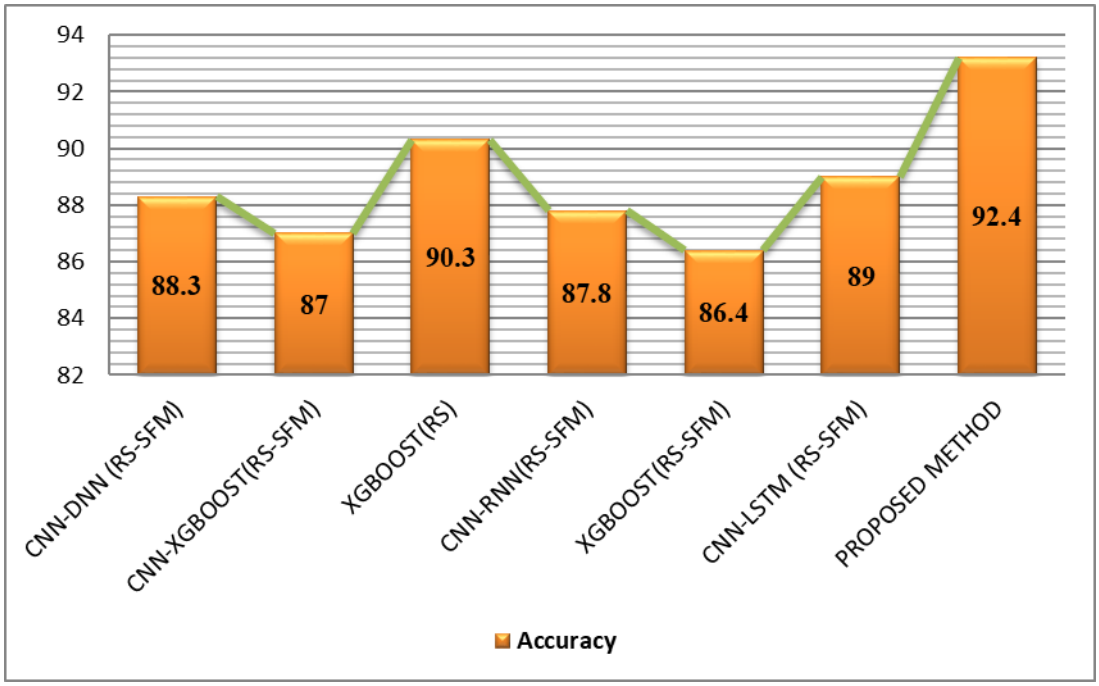

Figure 4: Model accuracy comparison between BYPEM Model and existing approaches.

4. CONCLUSION

This research underscores the efficacy of ensemble learning techniques, with particular emphasis on XGBoost, in accurately predicting Broccoli yield (kg). The study involved thorough data preprocessing, including the handling of missing values, removal of outliers, and Z-score normalization, ensuring the models were built on clean, standardized data. Feature selection was employed to identify key predictors, including pH, Moisture_Content, Nitrogen_Content, and Temperature, which significantly contributed to the enhancement of model accuracy.Among the models assessed, XGBoost demonstrated the highest performance across all evaluation metrics. Specifically, it achieved a Mean Absolute Error (MAE) of 0.17, a Mean Squared Error (MSE) of 0.10, a Root Mean Squared Error (RMSE) of 0.26, and an R^2 score of 0.91, indicating that the model explains 93.5% of the variance in Broccoli yield. Additionally, XGBoost attained 92.4% accuracy, further reinforcing its precision in predicting yield. In comparison, the Random Forest Regressor and Extra Trees Regressor models exhibited higher error rates and lower R^2 values, thereby highlighting XGBoost's superior predictive capability. This study emphasizes the critical role of ensemble learning, particularly boosting methods, in improving the accuracy and reliability of agricultural yield predictions. Future research could benefit from the integration of additional features, such as environmental and sensor data, to further enhance prediction accuracy and extend the applicability of the model to other crops and regions.

REFERENCES

[1] Ollio, I., Santás-Miguel, V., Gómez, D. S., Lloret, E., Sánchez-Navarro, V., Martínez-Martínez, S., ... & Zornoza, R. (2023). Effect of biofertilizers on broccoli yield and soil quality indicators. *Horticulturae*, *10*(1), 42.

[2] Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Hybrid deep learning-based models for crop yield prediction. *Applied artificial intelligence*, *36*(1), 2031822.

[3] Lohachov, M., Korei, R., Oki, K., Yoshida, K., Azechi, I., Salem, S. I., & Utsumi, N. (2024). RNN-Based Approach for Broccoli Harvest Time Forecast. *Agronomy*, *14*(2), 361.

Research Article

[4] Takahashi, M., Nakano, Y., & Sasaki, H. (2018). Increasing the yield of broccoli (Brassica oleracea L. var. italica) cultivar 'Yumehibiki'during the off-crop season by limiting the number of lateral branches. *The Horticulture Journal*, *87*(4), 508-515.

[5] Farias, E. D. S., Sant'ana, L. C. D. S., Melo, J. B., Santana, P. A., & Picanço, M. C. (2021). Impact of Diamondback Moth Density and Infestation Timing on Broccoli Yield. *Neotropical Entomology*, *50*, 298-302.

[6] Kim, S., Kim, S., Kiniry, J. R., & Ku, K. M. (2021). A hybrid decision tool for optimizing broccoli production in a changing climate. *Horticulture, Environment, and Biotechnology*, *62*, 299-312.

[7] Cammarano, D., Taylor, M. A., Thompson, J. A., Wright, G., Faichney, A., Haacker, R., ... & White, P. J. (2020). Predicting dates of head initiation and yields of broccoli crops grown throughout Scotland. *European Journal of Agronomy*, *116*, 126055.

[8] Patra, S. K., Poddar, R., Pramanik, S., Gaber, A., & Hossain, A. (2022). Crop and water productivity and profitability of broccoli (Brassica oleracea L. var. italica) under gravity drip irrigation with mulching condition in a humid sub-tropical climate. *Plos one*, *17*(3), e0265439.

[9] Luna, J. M., Sullivan, D., Garrett, A. M., & Xue, L. (2020). Cover crop nitrogen contribution to organic broccoli production. *Renewable Agriculture and Food Systems*, *35*(1), 49-58.

[10] Zhou, C., Hu, J., Xu, Z., Yue, J., Ye, H., & Yang, G. (2020). A monitoring system for the segmentation and grading of broccoli head based on deep learning and neural networks. Frontiers in plant science, 11, 402.

[11] You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017, February). Deep gaussian process for crop yield prediction based on remote sensing data. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

[12] ICAR Reseach Data Repository For Knowledge Management.

[13] Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and electronics in agriculture, 177, 105709.

[14] Ciancaleoni, S., Onofri, A., Torricelli, R., & Negri, V. (2016). Broccoli yield response to environmental factors in sustainable agriculture. *European Journal of Agronomy*, *72*, 1-9.

[15] Kim, S., Kim, S., Kiniry, J. R., & Ku, K. M. (2021). A hybrid decision tool for optimizing broccoli production in a changing climate. *Horticulture, Environment, and Biotechnology*, *62*, 299-312.

[16] Johansen, T. J., Mølmann, J. A., Bengtsson, G. B., Schreiner, M., Velasco, P., Hykkerud, A. L., ... & Seljåsen, R. (2017). Temperature and light conditions at different latitudes affect sensory quality of broccoli florets (Brassica oleracea L. var. italica). *Journal of the Science of Food and Agriculture*, *97*(11), 3500-3508.

[17] Šola, I., Gmižić, D., Pinterić, M., Tot, A., & Ludwig-Müller, J. (2024). Adjustments of the phytochemical profile of broccoli to low and high growing temperatures: Implications for the bioactivity of its extracts. *International journal of molecular sciences*, *25*(7), 3677.

[18] Scuderi, A., Timpanaro, G., Branca, F., & Cammarata, M. (2023). Economic and environmental sustainability assessment of an innovative organic broccoli production pattern. *Agronomy*, *13*(3), 624.

[19] Wang, H., He, Y., Zhang, W., Liao, J., & Zheng, Q. (2024). Integration of Cultivation Techniques and Innovation of Production Models for Specialty Vegetable: Broccoli. *Journal of Modern Business and Economics*, *1*(3).

[20] Karso, B. A., yousif Yousif, R., Mustafa, H., & Mohammad, D. (2022). Inventorying the most common broccoli pest insect and assessing the effectiveness of sticky traps in reducing damage. *NTU Journal of Agriculture and Veterinary Science*, *2*(1), 5-8.