Content list available at ITC

Techno-Science Research Journal

Journal Homepage: http://techno-srj.itc.edu.kh/

Techno-Science Research Journal

The Forecasting for Cambodia’s Rice Production Using Multiple Linear Regression and Tree – Based Ensemble Learning Methods

Cambodia

Received: 17 June 2025; Revised: Accepted: xxxx; Available online: xxxx

Abstract: *Cambodia's rice sector has undergone significant transformation over the past decades, evolving into a cornerstone of the national economy. Despite improvements in cultivation techniques and land use, the sector remains vulnerable to climate variability, inefficient resource allocation, and limited data-driven planning. To address these challenges, this study presents a comprehensive prediction framework that leverages both traditional Multiple Linear Regression (MLR) and advanced tree-based ensemble learning methods such as Random Forest, Gradient Boosting, and XGBoost to predict Cambodia’s annual rice production based on rice and climatic data of Cambodia. A Cambodia comprehensive dataset, compiled from the Food and Agriculture Organization (FAO) and World Bank, was used to train and evaluate the models, incorporating variables such as harvested area, annual production record, precipitation, and surface temperature. The results reveal that XGBoost achieved the highest predictive accuracy ($R^2 = 0.9511$), while feature importance analysis highlights the dominant influence of harvested area on rice yield. This work marks one of the first applications of ensemble learning in Cambodia's agricultural predictions and offers a decision – support tool that aligns with the nation's Digital Economy and Industrial Development Policies. The findings also contribute to national strategies on food security, smart agriculture, and enhance the global Sustainable Development Goals (SDGs).*

Keywords: Rice Production Forecasting; Multiple Linear Regression; Ensemble Learning Methods; Predictive Analytics; Cambodia Agriculture; Sustainable Development Goals; Machine Learning

1. INTRODUCTION[1]

1.1 Overview

Rice is Cambodia’s most crucial agricultural commodity, serving as the primary livelihood for millions and contributing approximately 60% to the national agricultural GDP, and represent as a major export agricultural product of the country [1]. According to the report from World Bank in 2015, during 2004 and 2012, the agricultural sector growth, particularly in rice production, was instrumental in reducing the national poverty rate from 53% to 18% [2]. Recognizing its strategic importance, the Roal Government of Cambodia (RGC) has enacted several initiatives, including the “Policy paper on the Promotion of Paddy Production and Rice Export” [3] and the Cambodia Industrial Develoopment Policy [4].

To further modernize agriculture, Cambodia has adopted the “Digital Economy and Society Policy Framework 2021 – 2035”, which promote digital transformation in agriculture, integration of artificial intelligence (AI) and machine learning (ML), and the development of data – driven rural strategies [5]. These policies align with multiple Sustainable Development Goals (SDGs) [6], such as zero hunger (SDG2), climate resilience (SDG13), and innovation (SDG9) by emphasizing evidence – based planning and infrfstructure development.

1.2 Literature Review

Many studies have explored the use of statistical and machine learning techniques to predict rice production or yield in various regions. Traditional methods such as time series analysis and linear regression techniques have been commonly applied due to their simplicity interpretability [7]. Over the last decades, advanced machine learning approaches like artificial neural networks, support vector machine, and decision tree – based algorithms have gained attention of their ability to capture nonlinear relatioships and improve predictive accuracy [8], [9], [10]. Additionally, other recent studies also tried to

* Corresponding author:
E-mail:

compare the performance between linear regression techniques with advance models such as artificial neural networks [11] and ensemble models.

A recent study of Patrio et al in 2023 focused on the prediction of rice production in Sumatra island, aimed at compare linear regression with ensemble and tree – based model, resulted that linear regression performed better than others in term of R^2_{score}, MAE, and MSE [12]. This result was mainly due to the linearity relationship of the attributes. Another study by Wijayanti et al in 2024, investigates rice yield prediction using the the XGBoost algorithm, leveraging combined datasets from the FAO and World Bank aimed at enhancing agricultural forecasting and policy decisions through accurate machine learning modeling. The datasets include variabls such as rice yield, pesticide use, rainfall, and temperature. The result showed that XGBoost is highly effective for rice yield prediction when feature engineering and data integration are optimized [13]. Despite the increasing adoption of predictive analytics globally, Cambodia has lacked context-spcific, data – driven tools, and limited research of predictive analytics applied in agricultural fields. This study addresses these gaps by introducing the use of localized climate and agricultural data with the state-of-the-art ensemble models for the first time to forecast rice production, aiming to enhance policy development for climate-resilient agriculture (SDG 13), contribute to innovation and sustainable infrastructure in the agricultural sector (SDG 9), support the Cambodia Industrial Development Policy and to fostering the Digital Economy and Society Policy Framework of the country.

2. METHODOLOGY

2.1 Dataset

The dataset used in this study spans from 1961 to 2022 and was compiled from reputable sources, including the Food and Agriculture Organization (FAO) and the World Bank. It was supplemented with additional Cambodia – specific datasets to ensure contextual relevance. The final dataset of 62 observations with four key variables recorded each year, resulting in 248 data points in total.

- Production (Tons): Annual rice production in Cambodia.
- Harvested Area (Hectares): Total area of land harvested for rice.
- Mean Precipitation (mm): Average annual rainfall.
- Mean Surface Temperature (°C): Average annual surface temperature.

Table 1. Statistical summary of the dataset

Variable	Mean	Std	Min	Max
Production	4396868.10	3347409.38	538000.00	12207000.00
Harvested Area	2074386.60	726660.87	555000.00	3498000.00
Mean Precipitation	1846.01	156.02	1477.24	2215.91
Mean Surface Temperature	27.13	0.404	26.3	27.99

2.2 Data Preprocessing

Prior to model development, a series of preprocessing steps were undertaken to improve data quality and model performance interpretability:

- Handling Missing Values: Any missing or anomalous values were identified and treated using interpolation and imputation techniques such as backward and forward.
- Feature Transformation: To stabilize variance and improve normality, logarithmic and square transformations were applied to both independent and dependent variables where appropriate.

The mathematical formulations for transformations used include:

$$y' = log\,(x) \qquad \text{(Eq. 1)}$$

Where:

$y' =$ the transformed values after applying natural logarithm
$x =$ the original vavues before applying natural logarithm

$$y' = x^2 \qquad \text{(Eq. 2)}$$

Where:

$y' =$ the transformed values after applying square transformation
$x =$ the original values before applying square transformation

These transformations aimed to enhance the linear relationship between predictors and the target variable.

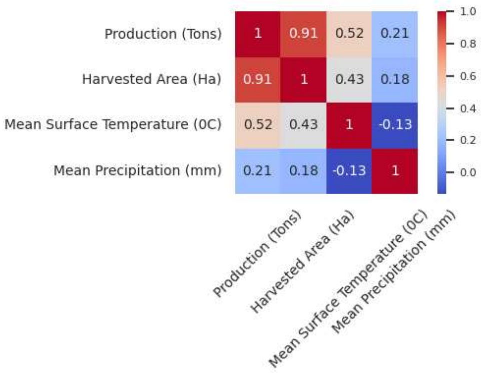

Fig. 1. Correlation Heatmap of the variables

2.3 Machine Learning Models

Multiple linear regression and tree-based ensemble learning (Gradient Boosting, XGBoost, and Random Forest) were implemented. Each model's theoretical foundation and application for regression tasks were outlined.

2.3.1 Multiple Linear Regression

Linear regression ia a statistical model, also known as a traditional approach that can solve the task by considering the historical dataset or parameter. There are two types of linear regression. When the analysis use only one independent variable to predict the dependent variable, it is considered as the simple linear regression (SLR). Otherwise, when the analysis use two or more independent variables, it is considered as the multiple linear regression (MLR). The mathematical representation of the MLR is written as;

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \qquad \text{(Eq. 3)}$$

Where:

$\hat{y}$ = dependent variable
β_0 = constant or $y_{intercept}$
$\beta_1, \beta_2, \ldots, \beta_n$ = slope coefficient
$X_1, X_2, \ldots, X_n$ = independent variable

Multiple linear regression is capable of analyzing one variable based on others various variables effectively. The slope shows the magnitude and direction of relationship between an independent variable and the dependent variable. It tells how much the dependent variable Y is expected to change when a specific independent variable change in one unit. Similarly, correlation coefficients are used to measure the degree of linear association between two variables, they show how strongly and in what direction the variables move together [14]. The correlation coefficient formula is written as;

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \qquad \text{(Eq. 4)}$$

The coefficient measures how well data can be represented in the regression line and can defines strength and direction of relationship between the dependent and independent variables.

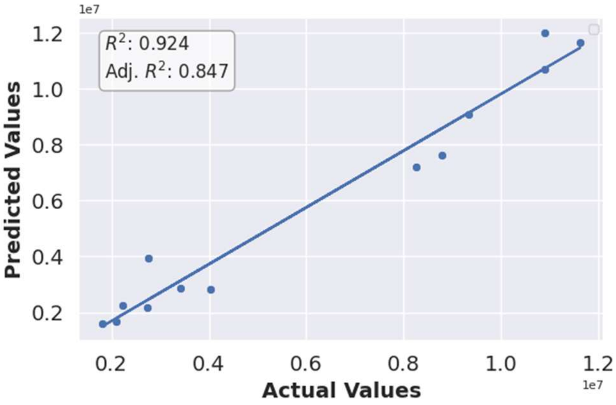

Fig. 2. Actual vs. predicted values for MLR model

2.3.2 Gradient Boosting

Gradient boosting is one of tree-based learning model, widely used for regression and classification work [15]. Gradient boosting regression builds multiple weak learners sequentially, optimizing model performance by minimizing a given loss function, such as mean squred error (MSE). At each step, the model tries to minimize the error (rediduals) of the previous model by fitting a new model to the gradient and improve model accuracy by learning from previous mistakes. Let y be the true value, and we want to predict y using a function F(x). The initial prediction model is written as;

$$F_0(x) = argmin\gamma \sum_{i=1}^{n} L(y_i, \gamma) \qquad \text{(Eq. 5)}$$

Where:

$F_0(x)$ = the initial model
$L(y_i, \gamma)$ = the loss function (mean squared error)
y_i = the actual target value

Compute the negative gradient (residuals) for each iteration $m = 1$ to M;

$$r_i^m = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \qquad \text{(Eq. 6)}$$

Fit a weak learner function $h_m(X)$, and find the best gradient decent step-size γ_m

$$\gamma_m = argmin_{\{\gamma\}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \qquad \text{(Eq. 7)}$$

Update the model;

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \qquad \text{(Eq. 8)}$$

Where:

$F_m(x)$ = the updated model (output)
$h_m(x)$ = the weak learner (decision tree)
γ_m = the learning rate or step size ($0 < \gamma_m \leq 1$)

The final model after M iterations can be written as;

$$F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \qquad \text{(Eq. 9)}$$

Hyperparameters was tuned using grid search with 10 – fold cross – validation, the tuned hyperparameters for Gradient Boosting shown in the table below.

Table 2. Hyperparameter tuning for gradient boosting

Search Range	Tuned Hyperparameters
N_estimators: [100, 200, 300, 500, 600], Learning_rate: [0.01, 0.05, 0.1, 0.2], Max_depth: [1, 2, 3, 4, 5, 6, ,7, 8, 9, 10], Sabsample: [0.5, 0.75, 1]	N_estimator: 100, Learning_rate: 0.05, Max_depth: 9, Sunsample: 0.5.

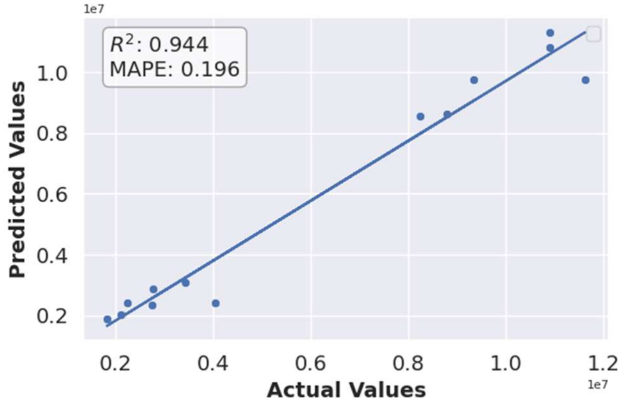

Fig. 3. Actual vs. predicted values for Gradient Boosting model

2.3.3 Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting) is an optimized implementation of Gradient Boosting algorithm, initially developed by Tianqi Chen as part of Distributed Machine Learning Community. Boosting tree algorithms based on the decision tree, which is known as classification and regression (CART). For regression tasks, CART divides the dataset into two subsets at each level according to the boundary for one variable until reaching the tree's maximun depth set by users [16]. It aims at extremely fast, scalable, and portable. XGBoost is used for supervised learning problems, where we use the training data (with multiple features) x_i to predict a target variable y_i. XGBoost extends traditional gradient boosting by including regularization elements in the objective function, improve generalization and prevents overfitting. The mathematical equation of the model is defined as;

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \qquad \text{(Eq. 10)}$$

Where:

$\hat{y}_i$ = the final predicted value for i^{th} data point
K = the number of trees in the ensemble
$f_k(x_i)$ = the prediction of K^{th} tree for the i^{th} data point

The objective function in XGBoost consists of two parts, a loss function and a regularization term. The loss function measures how well the model fits the data and the regularization term simplify complex trees. The general form of loss function is written as;

$$obj(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad \text{(Eq. 11)}$$

Where:

$l(y_i, \hat{y}_i)$ = the loss function which computes the difference between the true value
$\Omega(f_k)$ = the regularization term which discourages overly complex trees

By minimizing the model iteratively (instead of fitting the model all at once), starting with an initial prediction $\hat{y}_i^{\,0} = 0$ and add a new tree in each step to improve the model, then the updated predictions after adding the i^{th} tree can be written as;

$$\hat{y}_i^{\,t} = \hat{y}_i^{\,t-1} + f_t(x_i) \qquad \text{(Eq. 12)}$$

Where:

$\hat{y}_i^{\,t-1}$ = the prediction from the previous iteration
$f_t(x_i)$ = the prediction of the i^{th} tree for the i^{th} data point

The regularization term simplifies complex trees by penalizing the number of leaves in the tree and the size of leaf. It can be defined as;

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \qquad \text{(Eq. 13)}$$

Where:

T = the number of leaves in the tree
γ = the regularization parameter that controls the complexity of the tree
λ = the parameter that penalizes the squared weight of the leaves w_j

After splitting nodes in the trees, the informative gain for every possible split is computed. By calculating the information gain for every possible split at each node, XGBoost selects the split that results in the largest gain which effectively reduce the errors and improves the model's performance. The equation for information gain calculation is denoted as;

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} + \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \qquad \text{(Eq. 14)}$$

Where:

G_L, G_R = the sums of gradients in the left and right child nodes
H_L, H_R = the sums of Hessians in the left and right child nodes

Hyperparameters was tuned using grid search with 10 – fold cross – validation, the tuned hyperparameters for XGBoost shown in the table below.

Table 3. Hyperparameter tuning for XGBoost

Search Range	Tuned Hyperparameters
N_estimators: [100, 200, 300, 500, 600], Learning_rate: [0.01, 0.05, 0.1, 0.2], Max_depth: [1, 2, 3, 4, 5, 6, ,7, 8, 9, 10], Sabsample: [0.5, 0.75, 1]	N_estimator: 100, Learning_rate: 0.05, Max_depth: 3, Sunsample: 0.5.

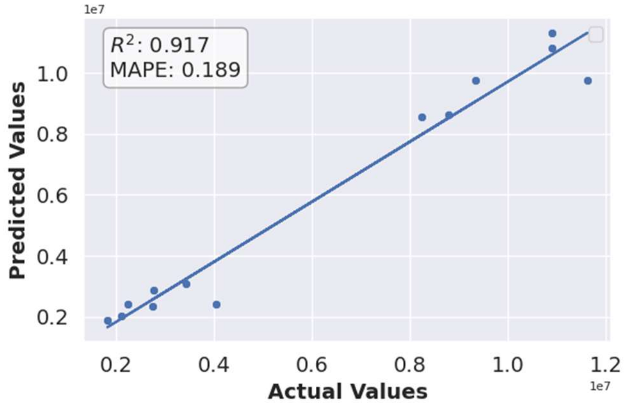

Fig. 4. Actual vs. predicted values for XGBoost model

2.3.4 Random Forest

Random forests are a supervised learning algorithm, also known as an ensemble learning method that is widely used in regression and classification problems. A random forest is essentially a collection of decision trees, where each tree is slightly different from the others [17]. It provides an improvement over bagged trees by way of a small tweak that decorrelates the trees. The number of decision trees are built on bootstrapped training samples and when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. A random forest built using quantitative variables, give output values also in quantitative values. It is assumed training data is independently selected from the original dataset [18]. A random forest uses Bootstrap Aggregation (Bagging) of a given dataset D with N observations.

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \qquad \text{(Eq. 15)}$$

Where:
x_i = the feature vector of the i^{th} sample
y_i = the target variable

By generating B bootstrap samples D_b by randomly selecting N with replacement $D_1, D_2, D_3, \ldots, D_B$. Each dataset D_b trains an individual regression tree denoted by $f_b(x)$. Each decision tree is trained independently on a bootstrap sample D_b. The trees are bult using the CART (Classification and regression Trees) algorithm. When building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. Each node in a decision tree split based on a feature x_j that minimizes a loss function. For regression, the best split is chosen by minimizing the Mean Squared Error (MSE).

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad \text{(Eq. 16)}$$

For a given feature x_j and split point s, we partition the data into two groups,

- Left node: $R_L = \{(x_i, y_i) | x_{ij} \le s\}$
- Right node: $R_R = \{(x_i, y_i) | x_{ij} > s\}$

The optimal split is determined by minimizing the weighted sum of MSEs;

$$argmin_{j,s}\left(\frac{|R_L|}{|R|}\sum_{i\epsilon R_L}(y_i - \bar{y}_L)^2 + \frac{|R_R|}{|R|}\sum_{i\epsilon R_R}(y_i - \bar{y}_R)^2\right) \qquad \text{(Eq. 17)}$$

$\bar{y}_L$ and $\bar{y}_R$ are the mean target values in the left and right nodes, respectively. Once all B decision trees are trained, each tree makes an individual prediction for a new input x as;

$$\hat{y}_b = f_b(x), for\ b = 1,2,3, \ldots, B \qquad \text{(Eq. 18)}$$

Therefore, the final prediction is the average of all tree inputs. The equation is written as;

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \qquad \text{(Eq. 19)}$$

Instead of using all features at each node, the random forest is built using $m = \sqrt{p}$ by selecting a random subset of features (known as m features out of total p features).

Hyperparameters was tuned using grid search with 10 – fold cross – validation, the tuned hyperparameters for random forest shown in the table below.

Table 3. Hyperparameter tuning for random forest

Search Range	Tuned Hyperparameters
N_estimators: [100, 200, 300], Max_depth: [none, 10, 20, 30], Min_samples_split: [2, 5, 10], Min_sample_leaf: [1, 2, 3], Max_features; [“auto”, “sqrt”]	N_estimators: 200, Max_depth: 20, Min_samples_split: 2, Min_sample_leaf: 1, Max_features; “sqrt”

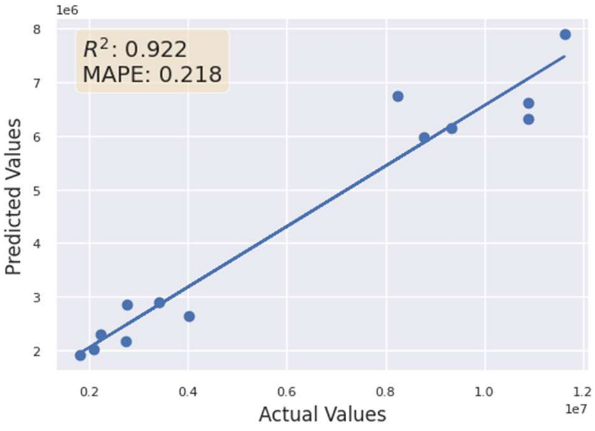

Fig. 5. Actual vs. predicted values for random forest model

2.4 Evaluation Metrics

In order to determine the forecast errors of the yearly rice production, different between the actual values and those predicted by the models were used. Determining the accuracy of the forecasts was indicated by calculating the values of the forecasting properties of the models. Coefficient of determination or known as R^2 score is the most used metric to measure the variance in the dependent variable that explained by the model, the higher R^2 score the better model performance [19].

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2} \qquad \text{(Eq. 20)}$$

Mean absolute error (MAE) is regularly employed in model evaluation studies [20]. The metric measure the average magnitude of errors between predicted and actual values, providing a straightforward interpretation of prediction accuracy.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \qquad \text{(Eq. 21)}$$

Mean absolute percentage error (MAPE) is a well known metric has been used for regression model [21]. It measures the accuracy of forecasting and regression models by expressing prediction errors as a percentage, offering an intuitive measure of how cose forecasts are to actual outcomes.

$$MAPE = \frac{100}{m}\sum_{i=1}^{m}\left|\frac{(y_i - \hat{y}_i}{y_i}\right| \qquad \text{(Eq. 22)}$$

3. RESULTS AND DISCUSSION

To evaluate and compare the models’ performance, three metrics such as R^2 (coefficient of determination), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error) were calculated.

Table 2. Models Performance Comparition

Models	R^2_{score}	MAE	MAPE
MLR	0.9236	816489.6390	0.1974
Gradient Boosting	0.9441	**600286.8961**	0.1962
XGBoost	**0.9511**	630508.4375	**0.1811**
Random Forest	0.9218	874510.8325	0.1839

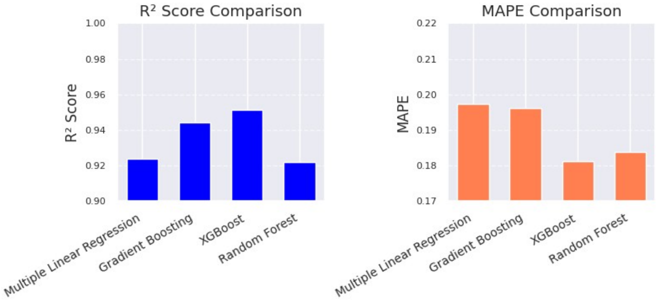

Fig. 6. Comparison of model performance

- XGBoost achieved the highest R² score (0.9511), indicating that it explained over 95% of the variance in rice production using the input variables. Although its MAE was slightly higher than Gradient Boosting, it had the lowest MAPE (18.11%), making it the most reliable in terms of relative forecasting accuracy.
- Gradient Boosting performed nearly as well, with a slightly better MAE but a marginally higher MAPE than XGBoost. This suggests its forecasts are closer in absolute value but slightly less consistent across different scales.
- Multiple Linear Regression, while traditional and more interpretable, lagged behind in accuracy. Its performance confirms the limitations of linear models in capturing complex, nonlinear relationships present in real-world agricultural systems.

- Random Forest, despite its robustness and ensemble nature, underperformed compared to XGBoost and Gradient Boosting. This may be due to overfitting or limitations in handling time-dependent features without tuning.

The feature importance analysis reveals that "Harvested Area (Ha)" is the most influential predictor across all models, particularly dominating in XGBoost and Gradient Boosting. "Mean Surface Temperature" is most important in the Multiple Linear Regression model, suggesting a strong linear relationship, while its impact is less significant in tree-based models. Transformed features, especially "Harvested Area transformed" and "Mean Temperature transformed", also contribute notably in Random Forest and Gradient Boosting, indicating the benefit of capturing non-linear patterns. In contrast, both original and transformed precipitation-related features show minimal importance, implying a limited role in the prediction of the target variable.

Accurate prediction of rice production enables proactive policy formulation, efficient resource allocation, and early response to climate risks. The results of this study suggest that integrating Random Forest-based predictive systems into national agricultural databases could significantly enhance strategic planning at both the governmental and farmer levels.

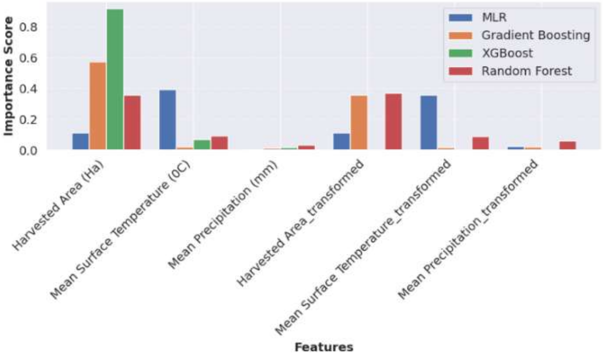

Fig. 7. Feature importance comparison

Furthermore, the findings support the policy directions outlined in Cambodia's Digital Economy and Society Policy Framework and contribute to the broader goals of food security, economic resilience, and climate-smart agriculture.

4. CONCLUSIONS

4.1. Conclusion

This study demonstrated the effectiveness of machine learning, particularly XGBoost in forecasting rice production using climate and agricultural data specific to Cambodia. Among the four models tested, XGBoost produced the highest predictive accuracy with an R^2 of 0.9511, confirming its ability to capture nonlinear and complex interactions among features such as harvested area, precipitation, and surface temperature.

Beyond model performance, this research holds strategic relevance to Cambodia's broader development trajectory. Accurate and data – driven agricultural forecasting supports national goals outlined in the Cambodia Industrial Development Policy (IDP) 2015 – 2025, which prioritizes modernization, digital transformation, and productivity growth in the agriculture sector. It also aligns with the Digital Economy and Society Policy Framework 2021–2035, which advocates for integrating artificial intelligence (AI), machine learning (ML), and data analytics into public planning and rural development.

The implementation of this predictive model directly contributes to multiple UN Sustainable Development Goals:

- SDG 2; Zero Hunger:

By improving the accuracy of rice production forecasts, this model helps reduce uncertainty in food supply, stabilize national food stocks, and support strategic interventions for food security.

- SDG 9; Industry, Innovation, and Infrastructure:

Table 3. Feature Importance Comparison

Features	MLR	Gradient Boosting	XGBoost	Random Forest
Harvested Area (Ha)	0.1120	**0.5736**	**0.9173**	**0.3566**
Mean Surface Temperature (°C)	**0.3918**	0.0212	0.0768	0.0912
Mean Precipitation (mm)	0.0029	0.0135	0.0149	0.0332
Harvested Area_transformed	0.1120	**0.3557**	0.0000	**0.3704**
Mean Surface Temperature_transformed	**0.3580**	0.0167	0.0000	0.0888
Mean Precipitation_transformed	0.0233	0.0191	0.0000	0.0603

The use of ensemble learning in agriculture reflects an innovative leap in data application, promoting smart infrastructure, digital agriculture systems, and technology-driven policy-making.

- SDG 13; Climate Action:

Incorporating climate variables (precipitation and temperature) into forecasting allows stakeholders to plan adaptively for changing weather patterns and support climate-resilient agricultural practices.

4.2. Future Work

To build upon the findings of this study, future research can explore:

- Expanded Feature Sets:

Include socio-economic indicators (e.g., input cost, labor availability), technological adoption levels, and policy shocks.

- Application Across Different Rice Growing Seasons:

Perforem predictive analysis using different data of each season in Cambodia and make prediction across different rice growing seasons.

ACKNOWLEDGMENTS

On the occasion of this achievement, I would like to extend my deepest gratitude to my advisor, Asst. Prof. Dr. PHAUK Sokkhey, for his expert guidaince, unwavering support, and invaluable insights throughout my academic journey. I would also like to express my sincere thanks to my co-advisor, Dr. HAS Sothea, for his consistent assistance and thoughtful advice during the implementation of my research. His mentorship has been instrumental in shaping my academic development, and this work would not have been possible without his encouragement and expertise.

REFERENCES

[1] S. Piseth, Y. Monyoudom, and H. Tynarath, “Cambodia’s Agri-Food Trade: Structure, New Emerging Potentials, Challenges & Impacts of Covid-19”.

[2] E. Asia and P. Region, *CAMBODIAN AGRICULTURE IN TRANSITION: OPPORTUNITIES AND RISKS*. 2015. [Online]. Available: https://documents1.worldbank.org/curated/en/805091467993504209/pdf/96308-ESW-KH-White-cover-P145838-PUBLIC-Cambodian-Agriculture-in-Transition.pdf

[3] “Policy Paper on The Promotion of Paddy Production and Rice Export,” Jul. 25, 2010. [Online]. Available: https://faolex.fao.org/docs/pdf/cam189808.pdf

[4] Royal Government of Cambodia, *Cambodia Industrial Development Policy*. Royal Government of Cambodia, 2015. [Online]. Available: https://cdc.gov.kh/wp-content/uploads/2022/04/IDP-English.pdf

[5] “CAMBODIA DIGITAL ECONOMY AND SOCIETY POLICY FRAMEWORK 2021 - 2035.” ROYAL GOVERNMENT OF CAMBODIA, May 2021.

[6] “The 2030 Agenda for Sustainable Development’s 17 Sustainable Development Goals (SDGs).” 4th SDG Youth Summer Camp – SDG Resource Document. [Online]. Available: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://sdgs.un.org/sites/default/files/2020-09/SDG%20Resource%20Document_Targets%20Overview.pdf

[7] P. Ly, L. S. Jensen, T. B. Bruun, and A. De Neergaard, “Factors explaining variability in rice yields in a rain-fed lowland rice ecosystem in Southern Cambodia,” *NJAS: Wageningen Journal of Life Sciences*, vol. 78, no. 1, pp. 129–137, Sep. 2016, doi: 10.1016/j.njas.2016.05.003.

[8] Erlin, A. Yunianta, L. A. Wulandhari, Y. Desnelita, N. Nasution, and Junadhi, “Enhancing Rice Production Prediction in Indonesia Using Advanced Machine Learning Models,” *IEEE Access*, vol. 12, pp. 151161–151177, 2024, doi: 10.1109/ACCESS.2024.3478738.

[9] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, “Rice crop yield prediction in India using support vector machines,” in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Khon Kaen: IEEE, Jul. 2016, pp. 1–5. doi: 10.1109/JCSSE.2016.7748856.

[10] N. Gandhi, O. Petkar, and L. J. Armstrong, “Rice crop yield prediction using artificial neural networks,” in *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, Chennai, India: IEEE, Jul. 2016, pp. 105–110. doi: 10.1109/TIAR.2016.7801222.

[11] M. Piekutowska *et al.*, “The Application of Multiple Linear Regression and Artificial Neural Network Models for Yield Prediction of Very Early Potato Cultivars before Harvest,” *Agronomy*, vol. 11, no. 5, p. 885, Apr. 2021, doi: 10.3390/agronomy11050885.

[12] U. Patrio, Y. Yuliska, and Y. Lulu Widyasari, “Predicting Rice Production In Sumatra Island Using Linear Regression,” in *Proceedings of the 11th International Applied Business and Engineering Conference, ABEC 2023, September 21st, 2023, Bengkalis, Riau, Indonesia*, Bengkalis, Indonesia: EAI, 2024. doi: 10.4108/eai.21-9-2023.2342997.

[13] E. B. Wijayanti, D. R. I. M. Setiadi, and B. H. Setyoko, “Dataset Analysis and Feature Characteristics to Predict Rice Production based on eXtreme Gradient Boosting,” *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 299–310, Feb. 2024, doi: 10.62411/jcta.10057.

[14] E. Sreehari and S. Srivastava, “Prediction of Climate Variable using Multiple Linear Regression,” in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India: IEEE, Dec. 2018, pp. 1–4. doi: 10.1109/CCAA.2018.8777452.

[15] M. Swamynathan, *Mastering Machine Learning with Python in Six Steps*. Berkeley, CA: Apress, 2017. doi: 10.1007/978-1-4842-2866-1.

[16] W. Dong, Y. Huang, B. Lehane, and G. Ma, “XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring,” *Automation in Construction*, vol. 114, p. 103155, Jun. 2020, doi: 10.1016/j.autcon.2020.103155.

[17] A. C. Mu, “Introduction to Machine Learning with Python”.

[18] J. K. Jaiswal and R. Samikannu, “Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression,” in *2017 World Congress on Computing and Communication Technologies (WCCCT)*, Tiruchirappalli, Tamil Nadu, India: IEEE, Feb. 2017, pp. 65–68. doi: 10.1109/WCCCT.2016.25.

[19] R. L. Sapra, “Using R2 with caution,” *Current Medicine Research and Practice*, vol. 4, no. 3, pp. 130–134, May 2014, doi: 10.1016/j.cmrp.2014.06.002.

[20] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature,” *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.

[21] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” *PeerJ Computer Science*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.