**International Journal of Informatics and Communication Technology (IJ-ICT)**
Vol. 14, No. 1, April 2025, pp. 164~173
ISSN: 2252-8776, DOI: 10.11591/ijict.v14i1.pp164-173

# Prediction of international rice production using long short-term memory and machine learning models

**Suraj Arya[1], Anju[1], Nor Azuana Ramli[2]**
[1]Department of Computer Science and Information Technology, Central University of Haryana, Mahendergarh, India
[2]Center for Mathematical Sciences, Universiti Malaysia Pahang AI-Sultan Abdullah, Kuantan, Malaysia

**Article Info**

***Article history:***

Received Sep 6, 2024
Revised Oct 9, 2024
Accepted Nov 19, 2024

***Keywords:***

Food security
Machine learning
Prediction
Regression
Rice production

## ABSTRACT

Rice, a staple food source globally, is in high demand and production across the world. Its consumption varies in different countries, with each nation having its unique way of incorporating rice into its diet. Recognizing the global nature of rice, its production is a crucial aspect of ensuring its availability, agriculture forecasting, economic stability, and food security. By predicting its production, we can develop a global plan for its production and stock, thereby preventing issues like famine. This paper proposes machine learning (ML) and deep learning (DL) models like linear regression, ridge regression, random forest (RF), adaptive boosting (AdaBoost), categorical boosting (CatBoost), extreme gradient boosting (XGBoost), gradient boosting, decision tree, and long short-term memory (LSTM) to predict international rice production. A total of nine ML and one DL models are trained and tested on the international dataset, which contains the rice production details of 192 countries over the last 62 years. Notably, linear regression and the LSTM algorithm predict rice production with the highest percentage of R-squared ($R^2$), 98.40% and 98.19%, respectively. These predictions and the developed models can play a vital role in resolving crop-related international problems, uniting the global agricultural community in a common cause.

***Corresponding Author:***

Nor Azuana Ramli
Center for Mathematical Sciences, Universiti Malaysia Pahang AI-Sultan Abdullah
26300, Kuantan, Pahang, Malaysia
Email: azuana@umpsa.edu.my

## 1. INTRODUCTION

In late 2023, a notable price surge and a global rice shortage surprised many. However, economic analysts had predicted this situation to occur as early as July 2023, when all non-Basmati rice in India would be subjected to export restrictions. Given that India produces nearly 40% of the world's rice, combined with the prolonged effects of climate change, Thailand's second-largest rice exporter also struggled to address this shortage. Before the full implementation of the export ban in July 2023, the market had already displayed signs of a rice shortage, evident through multiple price increases.

It is not easy to understand rice production since many factors influence it, such as climate change, political instability, pest attacks, crop diseases, changes in farming practices, and technology. According to the research by [1], temperature and rainfall fluctuations can significantly impact crop yields in Southeast Asia. Politically unstable situations often hinder rice cultivation, in addition to inconsistency in rules applicable or wars or trade embargoes under such situations, which can lead to significant interruption of agricultural activities. Another example is the ongoing conflict in Myanmar that has caused major disruptions to rice production and exports, resulting in global supply fluctuations and market price changes, affirming the industry’s susceptibility to geopolitical factors.

Accurate policy-making matters to the agriculture industry, for farmers, stakeholders, and policymakers. Hence, having an accurate prediction model will help correct food supply choices, resource allocation, and market strategies. Ensuring global food security and achieving sustainable development goal (SDG) 2: zero hunger, heavily relies on accurate rice production forecasts because rice is a staple to the majority of people across the globe. The importance of precise predictions on global food safety cannot be overemphasized. However, traditional crop yield prediction methods based on statistical models and historical data may not consider some of the complexities inherent in complex agricultural systems affected by climate change, soil conditions, pest invasion, and technological advancement.

Agricultural prediction has recently been highly improved by machine learning (ML) and deep learning (DL) techniques that have increased the accuracy of the prediction through training with large datasets on complex patterns. Some of these advanced methods include long short-term memory (LSTM), a recurrent neural network (RNN) type. These have successfully predicted time-series data because they can capture dependence over time and long-range correlation among the variables. Predicting crop yield is an example of how LSTM models have proved very effective because they are so good at capturing sequential patterns and temporal dependencies inherent in agricultural data.

In the last few years, a number of studies have shown the great potential of ML and DL models in agriculture yield prediction. A study done by [2] where incorporated convolutional LSTM, convolutional neural network (CNN), and hybridization of CNN and LSTM (CNN-LSTN) for predicting the annual rice yield at the county scale in Hubei Province, China. This research combined multiple sources of information such as gross primary productivity (GPP), ERA5 temperature (AT), soil-adapted vegetation index (SAVI), and MODIS remote sensing, which includes enhanced vegetation index (EVI), dummy spatial heterogeneity variable. These models have improved their prediction accuracy as soon as this dummy variable is introduced. It was found that incorporating spatial heterogeneity into models significantly improved prediction accuracy compared to remote sensing data alone. In addition, the ConvLSTM and CNN models outperformed the CNN-LSTM model.

Advanced models, such as CNN-LSTM-Attention models, have combined DL architectures and have been satisfactory in handling the nonlinear relationships within agricultural data, according to [3]. These models can handle massive, complex datasets, capturing the most important spatial and temporal variability and giving accurate predictions. Their results showed that advanced DL models considerably outperform traditional models, like random forest (RF) and extreme gradient boosting (XGBoost), which implies integrating these methods of effective multi-source data into crop yield prediction in the future. The result of this study can benefit policymakers and professionals working in the agriculture sector by making scientifically based policies to guide agricultural production for a safe and sustainable food supply.

This paper explores the work on rice production at an international level using LSTM and other machine-learning models. Therefore, the principal focus of this research will be to establish the effectiveness of these models in making predictions that are accurate and reliable for use in strategic planning and risk management in agriculture. Indeed, these advanced ML and DL models that leverage heterogeneous datasets will lead to high accuracy, outperforming traditional methods to provide new insights and tools for enhanced global food security. The contents of this paper are outlined as follows. Section 2 reviews related literature on cotton crop yield prediction using ML and DL techniques. Section 3 describes the methodology which involves data collection, determination of variables, data prepossessing, model design, model validation, and verification. Section 4 presents the findings and results of the experiment output, giving out some analysis with regard to checking the models. Finally, section 5 will discuss the findings and suggestions for further research.

## 2. LITERATURE REVIEW

Rice, the most widely consumed cereal grain globally, serves as a staple food for billions, particularly in Asia [4]. Its production is crucial for global food security and the livelihood of many farmers. Accurate prediction of rice production is not just vital, but a practical necessity for effective planning and decision-making in the agricultural sector [5]. The potential of ML and DL in predicting crop yield, including rice, is a promising area of research that has shown significant results in recent years [6].

A study conducted by [7] has significantly contributed to the field by using CNNs to predict rice yield. They utilized unmanned aerial vehicle (UAV) multispectral images and incorporated weather data at the heading stage. This innovative approach considers weather data in its analysis and adds valuable knowledge to the agricultural technology and remote sensing domain. The study's results demonstrate that a simple CNN feature extractor for UAV-based multispectral image input data can accurately predict crop yields. The models trained with weekly weather data performed the best. However, although the prediction accuracy was nearly the same, the spatial patterns of the predicted yield maps varied across different models. The study suggests that the robustness of within-field predictions should be evaluated alongside prediction accuracy.

Another study introduced a hybrid model, RaNN, which combines feature sampling and majority voting techniques from RF and multilayer Feedforward neural networks to predict crop yield [8]. The study was conducted in Punjab, India, the largest rice producer in the country. The dataset used in this study is robust, incorporating agriculture and weather datasets obtained from the Indian Meteorological Department Pune and Punjab Environment Information System (ENVIS) Center, Government of India. Results from this study revealed that RaNN produced an accurate model with minimal error, surpassing RF, multiple linear regression, support vector machine regression, decision tree, artificial neural network, boosting regression, and ensemble learner.

Several studies in agricultural research focus not only on methodology but also on the variables influencing model prediction accuracy [9]-[12]. For instance, [10] highlighted the significance of weather data and vegetation cover information in evaluating in-season rice yield estimation. They utilized the mobile app Canopeo and the conventional GreenSeeker handheld device to measure the normalized difference vegetation index (NDVI) during on-farm field experiments in rice-growing regions in 2018 and 2019. Additionally, they developed a generalized additive model (GAM) using log-transformed data for grain yield, including canopy cover and weather data during specific growth stages. However, the study's results were not as promising as anticipated, prompting the authors to emphasize the need for more field experiments to enhance the model's accuracy and robustness. In a different study, [9] collected real-time meteorological data and analysed the day-to-day impact of weather parameters on paddy cultivation. They proposed a robust optimized artificial neural network (ROANN) algorithm with genetic algorithm (GA) and multi-objective particle swarm optimization algorithm (MOPSO) to predict factors that could improve paddy yield. By optimizing input variables using GA and fine-tuning the neural network parameters, the proposed algorithm achieved maximum accuracy and minimum error rate.

Islam *et al.* [11] tackled the challenges of data quality, processing, and selecting suitable ML models with limited time-series data in a novel way. Their application of data processing techniques and a customized ML model significantly improved crop yield estimation accuracy at the district level in Nepal. Their finding that using remote sensing-derived NDVI alone was insufficient for accurate crop yield estimation, and that stacking multiple tree-based regression models together yielded better results, represents a significant advancement in the field. Finally, Bowden *et al.* [12] identified a relationship between monsoon variability and rice production in India, demonstrating the potential of RF modelling to reveal complex non-linearities and interactions between climate and rice production variability.

While most previous studies have used advanced techniques to predict rice production, our study takes a different approach. We focus on more straightforward ML and DL techniques, not only to prevent overfitting but also to make it easier for decision-makers to incorporate these models into their systems. Our global perspective, as opposed to a focus on a particular country, is a deliberate choice. We aim to provide new insights and tools that can be applied on a global scale, with the potential to significantly enhance food security worldwide.

The contributions of this work include: a) Global dataset: data utilized to forecast the rice production is related to a specific country India, Nepal. Proposed model has the details of 192 countries. Thus, its results have the international relevance. b) UAV and multispectral images: to predict the rice yield previous studies combining the weather data with multispectral images captured through UAV. Some papers are based on spatial patterns and yield maps. c) Dataset duration: dataset used for predictions contains the details of 65 years old rice production. This makes innovative use of historical dataset. d) Algorithms: background Studies are using the CNN, hybrid model, GA, swarm optimization algorithm and RaNN models to predict the rice production. Proposed models are the being an original contribution using time series model ARIMA and LSTM with the higher accurate results. e) International applicability: proposed study ensures availability of rice at the international level.

## 3. RESEARCH METHOD

This section delves into the rice production dataset and the methods used for predicting rice production. Python, a versatile and powerful programming tool, is the cornerstone of our model development. The following procedures are adopted to predict international rice production: i) data collection, ii) Identification of variables, iii) data pre-processing, iv) feature selection, v) data partitioning, vi) model training, and vii) model performance evaluation. Variable identification, such as weather conditions, soil quality, and previous year's production, data collection, and pre-processing are some of the most essential steps in training ML models. The model's effectiveness depends on the data's quality, consistency, and correctness. Figure 1 depicts the general flow of the process. The process began with data collection and so forth. The subsequent sub-sections will provide a detailed explanation of the process.

### 3.1. Data collection

This study used the international data set, which contains the details of rice production in 192 countries. This dataset is available on the open-access data repository ourworldindata.org [13]. It contains 10,128 rows and 4 columns.

### 3.2. Identification of variables

Our research has meticulously identified the variables crucial for predicting rice production. The entity under discussion, a significant contributing factor affecting rice production, has been carefully considered. We have also identified the regions that play a key role in this analysis. Our approach, which includes considering global entities, instils confidence in the accuracy of our predictions.

### 3.3. Data pre-processing

Data pre-processing involves preparing the data for the ML model. This includes taking necessary actions to improve its usability and ensure its proper format and structure, such as handling missing values, data inconsistencies, and conflicts. The rice production dataset initially contains a total of 10,128 rows and 4 columns. After removing the countries with discontinuous values from 1961 to 2022, the dataset contains 9,300 rows and 4 columns. The top five rows of the rice production dataset are displayed in Table 1.

### 3.4. Feature selection

Feature selection, the next crucial step in the pipeline, involves identifying and reducing the dataset to the most significant features. This process also involves removing features that do not affect the output variable. In our case, the ‘code’ feature is the selected feature.

### 3.5. Data partitioning

In this phase, the dataset was divided into two parts: training and testing for the ML and DL models. This study selected a commonly used ratio for balancing training and testing, 80:20.

### 3.6. Model training

Training data is provided as input to all of the ML and DL models. This paper applied the following models: linear regression (LR), RF regressor (RFR), XGBoost regressor, decision tree regressor (DTR), AdaBoost regressor (ABR), gradient boosting regressor (GBR), CatBoost regressor, ridge regressor (RR), and LSTM. The above-mentioned ML and DL models were selected because these models provide the best R-squared ($R^2$) values based on previous studies. We have utilized various regression models for making predictions since our dataset did not exhibit any time series properties.

#### 3.6.1. Linear regression

LR, a type of supervised ML regressor algorithm, is characterized by its interpretability. It comes in two types: simple linear regression with just one independent variable and multiple linear regression with more than one independent variable [14]. The equation for simple linear regression is as (1):

$$y = \beta_0 + \beta_1 x \tag{1}$$

where $y$ is dependent variable, $x$ is independent variable, $\beta_0$ is intercept, and $\beta_1$ is slope. Meanwhile, the equation for multiple regression:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_n x_n + \epsilon \tag{2}$$

where, for $i$ = $n$ observations: $y_i$ is dependent variable, $x_1$, $x_2$,.., $x_n$ are the independent variables, $\beta_0$ is $y$-intercept/constant, and $\beta_n$ is slope coefficients for each independent variable [15].

#### 3.6.2. Random forest regressor

RFR is a type of ensemble learning that improves accuracy by combining multiple decision trees as shown in Figure 2. It is used for classification and regression problems. It uses the ensemble’s bagging, boosting, and stacking methods for random feature selection [16], [17]. Different parameters were used to tune this algorithm. Some of the most widely used are:
max_depth: it indicates the maximum depth of each decision tree used in this model.
n_estimators: this parameter indicates the number of decision trees this model will use.

Table 1. Sample of international rice production dataset

| Index | Entity | Code | Year | Rice \| 00000027 \|\| production \| 005510 \|\| tonnes |
|---|---|---|---|---|
| 0 | Afghanistan | AFG | 1961 | 319000.0 |
| 1 | Afghanistan | AFG | 1962 | 319000.0 |
| 2 | Afghanistan | AFG | 1963 | 319000.0 |
| 3 | Afghanistan | AFG | 1964 | 380000.0 |
| 4 | Afghanistan | AFG | 1965 | 380000.0 |

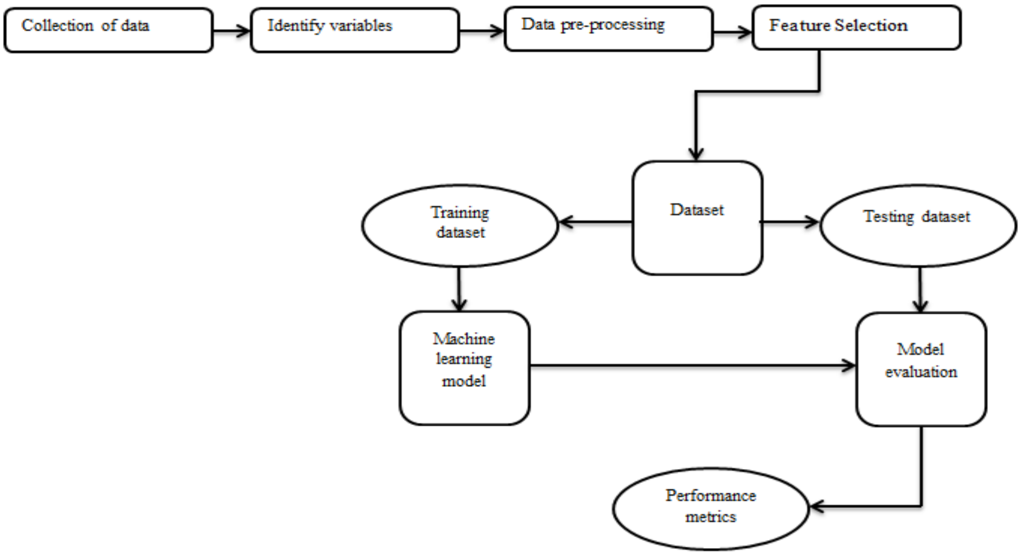

Figure 1. Flow of the proposed methodology

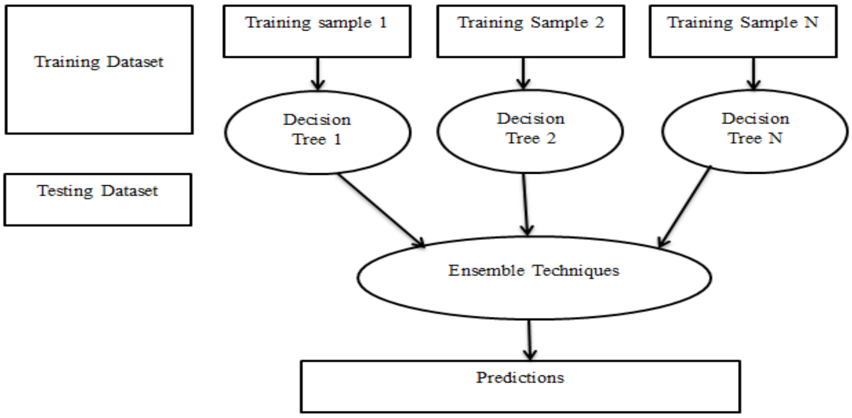

Figure 2. Random forest

### 3.6.3. XGBoost regressor

This regressor is the extension of the GBR. It is used to improve the performance of ML models with the help of objective functions. The objective function adopted by the XGB regressor is a mean squared error (MSE), but it can also be other functions like mean absolute error (MAE) or Huber loss. This regressor utilizes regularization techniques, such as L1 (lasso regularization) and L2 (ridge regularization), to prevent overfitting [18], [19].

### 3.6.4. AdaBoost regressor

AdaBoost is an ensemble method that utilizes the boosting technique and a decision tree as a base model. It is known as adaptive boosting because weights are reassigned to each instance, and higher weights are reassigned to incorrectly classified instances [20]. This model uses the learning rate and number of base models as parameters. AdaBoost has less of a possibility of overfitting than the other models, and it can be used to integrate other models to improve performance [21].

### 3.6.5. Long short-term memory

The LSTM is a well-known algorithm for estimating time-series and sequential datasets. LSTM can handle long-term dependencies to predicate the target variable. Other variants of LSTM include classic LSTM, stacked LSTM, and bidirectional LSTM. LSTM uses a memory cell to store information for long periods. It has three gates: input, forget, and output. The input gate determines what information is added in the cell state, the forget gate tells us which type of information is removed, and the output gate determines the output from the memory cell [22].

## 3.7. Model performance evaluation

All ML models and LSTM performance were evaluated based on the following metrics. The metrics are given below:

### 3.7.1. Mean squared error

MSE is the average of squared differences between actual and predicted values. It measures the average squared magnitude of errors [23]. The formula for MSE is given as:

$$\mathrm{MSE} = \sum_{k=1}^{n} \frac{(y_k - x_k)^2}{n} \quad (3)$$

### 3.7.2. Mean absolute error

MAE is the average of differences between actual and predicted values. It measures the average magnitude of errors [24]. The formula for MAE is given as:

$$\mathrm{MAE} = \sum_{k=1}^{n} \frac{|y_k - x_k|}{n} \quad (4)$$

### 3.7.3. Root mean squared error

Root mean squared error (RMSE) is the square root of MSE. It measures the standard deviation of residuals [25]. The formula for RMSE is given as:

$$\mathrm{RMSE} = \sqrt{\sum_{k=1}^{n} \frac{(y_k - x_k)^2}{n}} \quad (5)$$

### 3.7.4. R-squared

$R^2$, also known as a coefficient of determination, is a statistical technique that measures the goodness of fit of a regression model. The value lies between 0 and 1 [26].

$$\mathrm{R} - \mathrm{squared} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \quad (6)$$

# 4. RESULTS AND DISCUSSION

Before pre-processing data, it is best to analyse its statistical properties to gain insights. This includes examining statistical properties of the rice production dataset as tabulated in Table 2. The mean and standard deviation of the rice production dataset are 22800648.61642688 and 81591808.49779616, respectively. The mean of the dataset, which represents the average rice production, is a key statistical property to consider.

One DL and eight ML models were developed to predict the rice production of 192 countries. Of these, 42 countries have discontinuous and zero rice production. So, we have removed these countries and did not consider them for analysis. This dataset has following features as tabulated in Table 3. 'Entity', 'Code', 'Year', 'rice | 00000027 || production | 005510 || tonnes'.

Table 2. Statistical properties of rice dataset

| Index | Year | Rice \| 00000027 \|\| production \| 005510 \|\| tonnes |
|---|---|---|
| Count | 9300.0 | 9300.0 |
| Mean | 1991.5 | 22800648.61642688 |
| Std | 17.89649237104884 | 81591808.49779616 |
| Min | 1961.0 | 0.0 |
| 25% | 1976.0 | 37970.25 |
| 50% | 1991.5 | 403196.0 |
| 75% | 2007.0 | 4168674.0 |
| Max | 2022.0 | 789045300.0 |

Table 3. Sample of international rice production dataset after pre-processing

| Index | Entity | Code | Year | Rice \| 00000027 \|\| Production \| 005510 \|\| tonnes |
|---|---|---|---|---|
| 9295 | Zimbabwe | ZWE | 2018 | 1363.32 |
| 9296 | Zimbabwe | ZWE | 2019 | 1134.0 |
| 9297 | Zimbabwe | ZWE | 2020 | 750.0 |
| 9298 | Zimbabwe | ZWE | 2021 | 2908.0 |
| 9299 | Zimbabwe | ZWE | 2022 | 1923.32 |

Figure 3 displays a bubble plot of the top ten entities' average production (measured in 1000 tonnes) from 2017 to 2022. The chart includes the world, Asia, lower-middle-income countries, upper-middle-income countries, Southern Asia (FAO), Eastern Asia (FAO), China (FAO), India (FAO), and Southeastern Asia (FAO). Each entity is represented by a bubble, with the size correlating to the magnitude of rice production. Larger bubbles indicate higher production levels. From the plot, it can be seen that China and India have the largest bubbles, indicating they are the top producers. Asia as a region also shows significant production, reflecting the combined output of its countries. In contrast, lower-middle-income countries have notable production levels, highlighting their contribution to global rice production. This plot compares rice production effectively across different regions and income groups, providing insights into agricultural trends, and economic factors related to rice cultivation.

Table 4 displays the build and test time taken by the algorithms. The RF algorithm has the longest build time compared to the other models. On the other hand, LSTM's testing time is longer than that of the other algorithms. Reuß *et al.* [27] indicated that while LSTM has superior performance, it necessitates higher computational effort, requiring GPUs or longer computation time. Table 5 demonstrates error measurements and $R^2$values of different ML and DL techniques. The linear regression algorithm has the slightly highest $R^2$, with 98.40%, among ML and DL models. This highest $R^2$ shows that the model data is perfectly fit for regression. Table 6 shows the predicted value of the top five rice production entities and five randomly selected countries in tonnes using linear regression. Various models were developed to predict rice production, but only the linear regression model and LSTM were used to predict the rice production of five entities based on their performance.

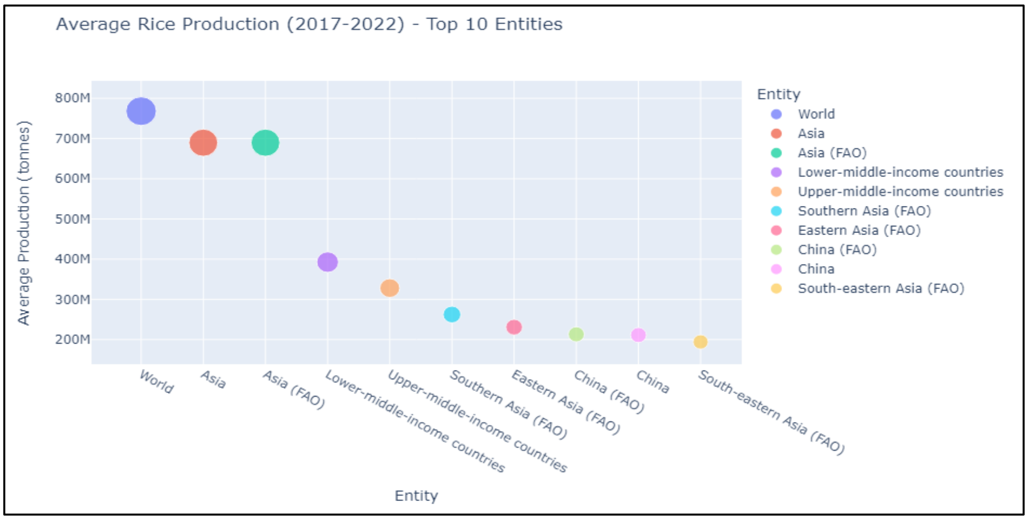

Figure 3. Average rice production from 2017 to 2022

Table 4. Build and test time for all models

| No | Algorithms | Build time (seconds) | Test time (seconds) |
|---|---|---|---|
| 1. | Random forest | 9.724 | 0.060 |
| 2. | AdaBoost | 0.281 | 0.003 |
| 3. | CatBoost | 4.439 | 0.013 |
| 4. | XGBoost | 0.250 | 0.005 |
| 5. | Gradient boosting | 3.600 | 0.002 |
| 6. | Decision tree | 0.151 | 0.001 |
| 7. | Linear regression | 0.012 | 0.001 |
| 8. | Ridge regression | 0.005 | 0.001 |
| 9. | LSTM | 2.584 | 0.274 |

Table 5. Performance of ML and DL models

| Algorithms | MAE | MSE | RMSE | R-squared |
|---|---|---|---|---|
| Random forest | **0.0021** | **0.0002** | 0.0166 | 0.9710 |
| AdaBoost | 0.0083 | 0.0005 | 0.0236 | 0.9417 |
| CatBoost | 0.0024 | **0.0004** | 0.0205 | 0.9561 |
| XGBoost | 0.0026 | 0.0005 | 0.0225 | 0.9472 |
| Gradient boosting | **0.0021** | **0.0001** | 0.0135 | **0.9831** |
| Decision tree | **0.0021** | **0.0002** | 0.0154 | 0.9752 |
| Linear regression | **0.0022** | **0.0004** | 0.0216 | **0.9840** |
| Ridge regression | 0.0025 | **0.0004** | 0.0219 | 0.9502 |
| LSTM | 0.0057 | **0.0001** | 0.0140 | **0.9819** |

Table 6. Predicted rice production of top five countries of ML and DL models

| | Predicted rice production linear regression | | | | | |
|---|---|---|---|---|---|---|
| Entity | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 |
| World | 831491200 | 841000200 | 850509200 | 860018200 | 869527200 | 879036200 |
| Asia | 750087100 | 758561900 | 767036800 | 775511700 | 783986500 | 792461400 |
| Asia (FAO) | 750087100 | 758561900 | 767036800 | 775511700 | 783986500 | 792461400 |
| Upper-middle-income countries | 376173600 | 380057700 | 383941700 | 387825800 | 391709900 | 395593900 |
| Lower-middle-income countries | 406570900 | 411991900 | 417412900 | 422833900 | 428254900 | 433675900 |
| | Predicted rice production using LSTM | | | | | |
| World | 825374400 | 836444800 | 847910976 | 859832448 | 871877632 | 885726080 |
| Asia | 751069440 | 762517248 | 775157504 | 788389184 | 802099712 | 818525760 |
| Asia (FAO) | 760280512 | 774286400 | 790011264 | 806535296 | 824263616 | 844865536 |
| Upper-middle-income countries | 343263776 | 344810208 | 346892000 | 348744768 | 350653376 | 353694400 |
| Lower-middle-income countries | 433157824 | 442021184 | 450856128 | 460013440 | 469265152 | 479333792 |
| | Predicted rice production using linear regression of random five countries | | | | | |
| France | 99096.89 | 99426.41 | 99755.94 | 100085.5 | 100415 | 100744.5 |
| India | 186767100 | 189072200 | 191377300 | 193682500 | 195987600 | 198292700 |
| China | 237836300 | 240090900 | 242345400 | 244599900 | 246854500 | 249109000 |
| Iran | 2730726 | 2757446 | 2784166 | 2810887 | 2837607 | 2864327 |
| Australia | 831291.6 | 837944.7 | 844597.8 | 851250.9 | 857904 | 864557.1 |

In 2025, Asia will produce 750,087,100 tonnes of rice. As different models have different predicted values, the average rice production for the year 2025 is 676,959,750 of the top five entities. The authors of this paper propose all these assumptions. Similarly, all predicted values of the top five entities are depicted in Table 6. Figure 4 demonstrates the difference in rice production from 1961 to 2022. All five countries had the maximum difference in rice production from the starting year (1961) to the ending year (2022). Japan showed the most significant change.

The two countries ( i.e., 69,210.00), in Western Europe and France, have the lowest maximum change in rice production. The percentage change from the predicted value in 2023 to the predicted value in 2030 for the upper-middle-income country is 0.14%. Similarly, we can calculate the percentage difference for any other country.

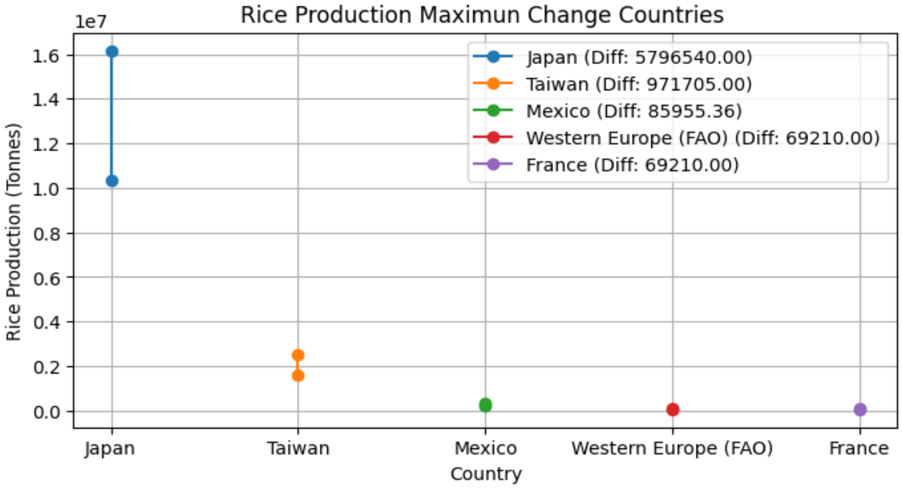

Figure 4. Rice production max changes countries

## 5. CONCLUSION

This paper predicts the international rice production of 150 countries with the help of ML and DL models, showcasing the potential of these innovative technologies in the field of agriculture. Nine models were developed, eight of which are ML, and only one is DL. The study concludes that the international average rice production will be 34,527,755.27 tonnes in 2025. In the Asia continent, the production of rice will be 750,087,100 in 2025. Trained ML and DL models can also predict the values for Asia (FAO), upper-middle-income, and lower-middle-income countries. Compared to the current rice production predicted value of upper-middle-income countries in the future, 0.14 % will increase internationally from 2023 to 2030. Thus, using these ML and DL models, we can predict the future value of rice production in 150 countries. The accuracy level of ML and DL models was measured using $R^2$. The linear regression model provides the best-predicted value of rice production compared to the other models. The $R^2$ value of this model is slightly the highest, showing goodness of fit. Thus, this paper can contribute to developing international agriculture strategies based on the outcome of this study. Nonetheless, a few limitations could be addressed in future research. The first limitation is the methodology, where this study employs nine models, but only one DL model is employed. A broader comparison involving diverse DL architectures could offer a more comprehensive performance evaluation. Another limitation is the assumption made during the development of the models. The study assumes that current socio-economic and policy conditions will remain constant, which may not be realistic in the real world. Changes in agricultural policies, trade agreements, or economic shocks could significantly impact production trends. Future studies should consider these variables to develop models better suited to real-world changes.

## ACKNOWLEDGEMENTS

Author thanks Research and Innovation Department, Universiti Malaysia Pahang Al-Sultan Abdullah for funding this paper under International Matching Grant with registration number UIC231524.

## REFERENCES

[1] B. T. Tan, P. S. Fam, R. B. R. Firdaus, M. L. Tan, and M. S. Gunaratne, “Impact of climate change on rice yield in Malaysia: a panel data analysis,” *Agriculture*, vol. 11, no. 6, p. 569, Jun. 2021, doi: 10.3390/agriculture11060569.

[2] S. Zhou, L. Xu, and N. Chen, “Rice yield prediction in Hubei Province based on deep learning and the effect of spatial heterogeneity,” *Remote Sensing*, vol. 15, no. 5, p. 1361, Feb. 2023, doi: 10.3390/rs15051361.

[3] J. Lu *et al.*, “Deep learning for multi-source data-driven crop yield prediction in Northeast China,” *Agriculture*, vol. 14, no. 6, p. 794, May 2024, doi: 10.3390/agriculture14060794.

[4] N. Annamalai and A. Johnson, “Analysis and forecasting of area under cultivation of rice in india: univariate time series approach,” *SN Computer Science*, vol. 4, no. 2, p. 193, Feb. 2023, doi: 10.1007/s42979-022-01604-0.

[5] 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)*, Jul. 2016, pp. 1–5, doi: 10.1109/JCSSE.2016.7748856.

[6] X. Han, F. Liu, X. He, and F. Ling, “Research on rice yield prediction model based on deep learning,” *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–9, Apr. 2022, doi: 10.1155/2022/1922561.

[7] M. S. Mia *et al.*, “Multimodal deep learning for rice yield prediction using UAV-based multispectral imagery and weather data,” *Remote Sensing*, vol. 15, no. 10, p. 2511, May 2023, doi: 10.3390/rs15102511.

[8] S. Lingwal, K. K. Bhatia, and M. Singh, “A novel machine learning approach for rice yield estimation,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 36, no. 3, pp. 337–356, Apr. 2024, doi: 10.1080/0952813X.2022.2062458.

[9] S. Muthukumaran, P. Geetha, and E. Ramaraj, “Multi-objective optimization with artificial neural network based robust paddy yield prediction model,” *Intelligent Automation & Soft Computing*, vol. 35, no. 1, pp. 215–230, 2023, doi: 10.32604/iasc.2023.027449.

[10] C. B. Onwuchekwa-Henry, F. V. Ogtrop, R. Roche, and D. K. Y. Tan, “Model for predicting rice yield from reflectance index and weather variables in lowland rice fields,” *Agriculture*, vol. 12, no. 2, p. 130, Jan. 2022, doi: 10.3390/agriculture12020130.

[11] M. D. Islam *et al.*, “Rapid rice yield estimation using integrated remote sensing and meteorological data and machine learning,” *Remote Sensing*, vol. 15, no. 9, p. 2374, Apr. 2023, doi: 10.3390/rs15092374.

[12] C. Bowden, T. Foster, and B. Parkes, “Identifying links between monsoon variability and rice production in India through machine learning,” *Scientific Reports*, vol. 13, no. 1, p. 2446, Feb. 2023, doi: 10.1038/s41598-023-27752-8.

[13] “Global food explorer,” *Our World in Data*, 2022. https://ourworldindata.org/explorers/global-food?Food=Rice&Metric=Production&Per+Capita=false (accessed Jul. 05, 2024).

[14] F. Rios-Avila and M. L. Maroto, “Moving beyond linear regression: implementing and interpreting quantile regression models with fixed effects,” *Sociological Methods & Research*, vol. 53, no. 2, pp. 639–682, May 2024, doi: 10.1177/00491241211036165.

[15] C. Y. Lim *et al.*, “Linearity assessment: deviation from linearity and residual of linear regression approaches,” *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 62, no. 10, pp. 1918–1927, Sep. 2024, doi: 10.1515/cclm-2023-1354.

[16] F. Özen, “Random forest regression for prediction of Covid-19 daily cases and deaths in Turkey,” *Heliyon*, vol. 10, no. 4, p. e25746, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25746.

[17] A. Harmayanti, I. P. Tama, F. Gapsari, Z. Akbar, and H. Juliano, “Cardiac biometrics and perceived workload regression analysis using random forest regressor in cognitive manufacturing tasks,” *International Journal of Mechanical Engineering Technologies and Applications*, vol. 5, no. 1, pp. 108–120, Jan. 2024, doi: 10.21776/MECHTA.2024.005.01.11.

[18] P. Panchal *et al.*, “XGBoost regression analysis of dielectric properties of epoxy resin with inorganic hybrid nanofillers,” *Journal of Macromolecular Science, Part B*, pp. 1–17, May 2024, doi: 10.1080/00222348.2024.2347746.

[19] H. Sharma, H. Harsora, and B. Ogunleye, "An optimal house price prediction algorithm: XGBoost," *Analytics*, vol. 3, no. 1, pp. 30–45, Jan. 2024, doi: 10.3390/analytics3010003.

[20] Y. Xv, Y. Sun, and Y. Zhang, "Prediction method for high-speed laser cladding coating quality based on random forest and AdaBoost regression analysis," *Materials*, vol. 17, no. 6, p. 1266, Mar. 2024, doi: 10.3390/ma17061266.

[21] S. S. Hussain and S. S. H. Zaidi, "AdaBoost ensemble approach with weak classifiers for gear fault diagnosis and prognosis in DC motors," *Applied Sciences*, vol. 14, no. 7, p. 3105, Apr. 2024, doi: 10.3390/app14073105.

[22] J. Wang, S. Hong, Y. Dong, Z. Li, and J. Hu, "Predicting stock market trends using LSTM networks: overcoming RNN limitations for improved financial forecasting," *Journal of Computer Science and Software Applications*, vol. 4, no. 3, pp. 1–7, 2024, doi: 10.5281/zenodo.12200708.

[23] V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Bridging Bayesian and minimax mean square error estimation via wasserstein distributionally robust optimization," *Mathematics of Operations Research*, vol. 48, no. 1, pp. 1–37, Feb. 2023, doi: 10.1287/moor.2021.1176.

[24] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/gmd-15-5481-2022.

[25] 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.

[26] C. Onyutha, "From R-squared to coefficient of model accuracy for assessing 'goodness-of-fits,'" *Geoscientific Model Development Discussions*, pp. 1–25, 2020.

[27] F. Reuß, I. Greimeister-Pfeil, M. Vreugdenhil, and W. Wagner, "Comparison of long short-term memory networks and random forest for sentinel-1 time series based large scale crop classification," *Remote Sensing*, vol. 13, no. 24, p. 5000, Dec. 2021, doi: 10.3390/rs13245000.

## BIOGRAPHIES OF AUTHORS

**Dr. Suraj Arya** is currently working as assistant professor in the Department of Computer Science and Information Technology and Deputy Director (Training and Placement) in Central University of Haryana, India. His academic qualifications are Ph.D. (Computer Science), M.Phil. (Computer Science) and M. Tech (Computer Science and Engineering). His research interests focus on machine learning (ML), internet of things (IoT), data warehousing and mining, system automation and patents writings. He has granted and files many patents. He has also published many research articles in international journals, book chapters and conferences. He can be contacted at email: surajarya@cuh.ac.in.

**Anju** is a research scholar of Central University of Haryana, India. She received her B.Tech. in Computer Science and Engineering from Maharshi Dayanand University Rohtak and M.Sc. in Computer Science from Chaudhary Bansi Lal University Bhiwani. She is currently doing her Ph.D. (Computer Science) from Central University of Haryana. Her research interests: ML, and IoT. She can be contacted at email: anju24sanga@gmail.com.

**Nor Azuana Ramli** is a senior lecturer in the Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah. She received her Ph.D. from Universiti Sains Malaysia, Master in Innovation and Engineering Design from Universiti Putra Malaysia and B.Sc. degree in Industrial Mathematics from Universiti Teknologi Malaysia. Her current research involves big data analytics, machine learning, deep learning, computer vision, data mining and artificial intelligence. She has published 57 research articles in reputed SCI and SCOPUS indexed journals and conferences. She can be contacted at email: azuana@umpsa.edu.my.