**JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY**
*https://e-journal.uum.edu.my/index.php/jict*

How to cite this article:
Mohamad Mohsin, M. F., Hassan, M. G., Mohd Sharif, K. I., Ismail, M. A., Hong, L. J., Umana, M. K., & Rahim@Abdul Rahim, H. (2025). Prediction of rice yields in a changing climate using the mobile rice yield prediction application. *Journal of Information and Communication Technology, 24*(2),1-18. https://doi.org/10.32890/jict2025.24.2.1

# Prediction of Rice Yields in a Changing Climate Using the Mobile Rice Yield Prediction Application

[1]Mohamad Farhan Mohamad Mohsin, [2]Mohamad Ghozali Hassan,
[3]Kamal Imran Mohd Sharif, [4]Mohd Azril Ismail, [5]Liew Jun Hong,
[6]Muhammad Khalifa Umana, & [7]Hairazi Rahim@Abdul Rahim
[1,5&6]School of Computing, Universiti Utara Malaysia, Malaysia
[2,3&4]School of Technology Management and Logistics,
Universiti Utara Malaysia, Malaysia
[7]Pusat Penyelidikan Sosio Ekonomi, Risikan Pasaran & Agribisnes,
Ibu Pejabat MARDI, Malaysia

*[1]farhan@uum.edu.my
[2]ghozali@uum.edu.my
[3]kamalimran@uum.edu.my
[4]azril@uum.edu.my
[5]junhong8074@gmail.com
[6]umana_m_khalifa@ahsgs.uum.edu.my
[7]hairazi@mardi.gov.my
*Corresponding author

Received: 2/10/2024 Revised: 13/1/2025 Accepted: 1/4/2025 Published: 30/4/2025

## ABSTRACT

In Malaysia, rice production is crucial for national food security but faces significant challenges due to unpredictable climate patterns and climate change. Traditional prediction methods often struggle to capture the complex interaction between climatic and agricultural factors that cause a critical gap in accurately forecasting rice yields. This paper examines the Mobile Rice Yield Prediction Application (MRYPA) through a comparative analysis of two regression models: A baseline multiple linear regression model and an enhanced version incorporating polynomial and interaction terms. A 10-year dataset from the Malaysian Meteorological Department (MetMalaysia) and the Department of Statistics Malaysia (DOA) were trained and tested on varying data splits and then evaluated using performance metrics such as R-squared, Mean Absolute Error, Mean Square Error, and Root Mean Square Error. The findings reveal that while the enhanced model demonstrates marginally better predictive accuracy in smaller datasets, both models exhibit comparable performance with larger training datasets. A 50% train-test split yielded the best results for both models, achieving an R-squared value of 0.9843 for the baseline model and marginal improvements for the enhanced model. These findings underscore the potential of integrating yield and climate data into user-friendly tools, such as MRYPA, to empower stakeholders to make informed, data-driven decisions for sustainable agriculture and national food security. Future work will refine the prediction accuracy and system utility by integrating different data sources and advanced machine-learning models.

**Keywords:** Agricultural forecasting, climate change, machine learning, rice yield prediction.

## INTRODUCTION

Rice is a vital food and the main dietary source for many Asian countries, including Malaysia. As stated in a report by Khazanah Research Institute (Omar et al., 2019), Malaysians consumed 87.9 kg of rice per person per year, which is significantly higher than the global average of 54.6 kg. This high consumption rate translates to approximately 2.7 million metric tonnes of national rice demand in 2016. Additionally, Malaysia had a net rice import of 776,448 metric tonnes in 2016, which corresponds to around 17.9 kg of imported rice per person — indicating a continued reliance on external supply to meet national demand. In 2021, Malaysia's paddy harvest surpassed 2.343 million metric tons. Despite this, it only met approximately 73% of the country's rice demand (Abidin & Dardak, 2023). As a result, Malaysia depends significantly on imports from Thailand, Vietnam, and Pakistan. Malaysia imported 821,869 metric tons of rice in 2016 and exported 45,421 tons, for a net import of 776,448 metric tons (Omar et al., 2019). In 2023, this import volume increased to over 1.22 million tons of various rice varieties (Abidin & Dardak, 2023).

In Malaysia, rice provides just 2.3% of Malaysia's gross domestic product (GDP); however, it plays a critical role in national food security. Any delays in supply can have serious social and economic effects. Therefore, ensuring a stable and sufficient rice supply is essential to maintain the national food supply and social stability. Achieving production stability is challenging due to climate change, fluctuating soil conditions, pest outbreaks, and unpredictable weather patterns (Herman et al., 2015). It restricts attempts to meet the increased demand for rice. As a result, this forces Malaysia to be more reliant on imports (Alam et al., 2013; Jabin et al., 2015).

Inconsistent weather patterns can affect rice production. The weather, such as extended dry spells, flash floods, and unusually high temperatures during key growth periods can impact harvests and pose a risk to national food security (Masud et al., 2014; Tan et al., 2021). One of the constraints in traditional prediction methods is the use of past averages and simple linear models, which struggle to account for the intricate connections between climate and agriculture (Tan et al., 2021). This shortcoming often leaves farmers and policymakers reacting to crises. It then hinders them from proactive resource management and long-term food planning.

Many current forecasting techniques assume a direct relationship between factors such as rainfall and harvest size. This is a simplification as the real-world relationship is much more complex. For instance, temperature and humidity often exhibit non-linear effects on rice yield. High temperatures may reduce yield when combined with low humidity. Meanwhile, moderate temperature increases under certain humidity levels can be beneficial. These non-linear relationships are poorly represented in traditional linear models, leading to suboptimal predictions. Incorporating polynomial features into the regression model can address these non-linearities by capturing quadratic or curvilinear effects. Besides that, the model's predictive capability can be enhanced through this integration (Schlenker & Roberts, 2009).

To address the limitations of existing prediction methods, we evaluate the predictive ability of the Mobile Rice Yield Prediction Application (MRYPA). This involves comparing a basic multiple linear regression model against an enhanced model that incorporates polynomial and interaction effects. Using datasets provided by MetMalaysia and the Department of Agriculture (DOA) Malaysia (2010-2021), we investigate whether these enhancements lead to improved prediction accuracy and a deeper understanding of the factors that influence rice harvesting. The output from this work can help the field of agricultural forecasting by creating a practical tool that combines climate and agricultural data to support proactive decision-making.

## RELATED WORK

### Rice Plantation in Malaysia

Rice is a product cultivated from paddy. It is considered the lifeblood of many countries, including Malaysia (Dorairaj & Govender, 2023). The rice industry ranks as the third most significant agricultural commodity in Malaysia after the rubber and oil palm sectors. As mentioned by Firdaus et al. (2020), rice provides a livelihood for over 322,830 farmers and contributes to safeguarding the nation's food supply. Although it does not fully meet domestic rice consumption, paddy production plays a crucial role in import substitution, thereby reducing reliance on foreign rice and stabilising market prices. The state of Kedah possesses the biggest area for paddy farming, accounting for almost 33% of the national total. Besides Kedah as a major rice producer, the states of Sarawak, Kelantan, and Perak also gain huge benefits from governmental programs aimed at creating "granary" zones equipped with specialised irrigation infrastructure and high-yield rice production (Firdaus et al., 2020).

Malaysia has several agencies that are responsible for rice plantation and production. The Ministry of Agriculture and Food Security (MOA) is Malaysia's leading agency for paddy management. Each agency related to paddy has different roles. The DOA implements agricultural policies and provides technical support. The Farmers' Organization Authority (FOA) focuses on the development of farmers. The Malaysian Agricultural Research and Development Institute, also known as MARDI, is tasked with boosting agricultural innovation. The agency that focuses on rural land development and rehabilitation, including rice plantation, is known as FELCRA or Federal Land Consolidation and Rehabilitation Authority.

Malaysian rice production has experienced a decreasing pattern over the last 10 years. The statistics have noted that the country's rice production has shrunk by over 11% since 2015. Among the factors that contribute to this scenario are land conversion, an ageing farmer population, and climate change. As a response to this declination, the Malaysian government has implemented subsidies and infrastructure development in designated agricultural zones (Makhtar et al., 2022; Mohd Ali et al., 2021; Siwar et al., 2014). Furthermore, the government also pushed for mechanisation and research into high-yield, disease-resistant paddy varieties, aiming to improve overall production and attract younger farmers (Baharudin & Waked, 2021). While the future remains uncertain, with climate change being a major concern, these efforts offer opportunities for revival. Ensuring a sustainable and productive paddy sector is vital for Malaysia's food security and cultural heritage.

## Rice Prediction Systems

There are two types of crop yield prediction. First is the process-based model, and second is the data-driven model (Chang et al., 2023). Process-based models simulate inputs for prediction, such as yield growth and evolution, based on weather, soil, and management factors. The two well-known process-based models are APSIM (Balboa et al., 2019) and DSSAT (Corbeels et al., 2016). The models can provide scientifically explainable results rooted in plant biology; nevertheless, their heavy reliance on experimental calibration and lack of adaptability to diverse environmental conditions limit their scalability and predictive performance. In contrast, data-driven models employ machine learning techniques. This approach excels in handling large datasets and can discover complex relationships between input variables and yielding outcomes. Several data-driven models, such as random forest regression, deep learning, and regression-based models, have demonstrated remarkable prediction accuracy. These models capture the combined effects of genetics, environment, and management practices to provide higher accuracy than process-based models. However, they often lack in terms of interpretability and generalizability as a reflection of the "black box" feature in machine learning models.

Several data-driven models have been developed for rice yield prediction. This includes neural networks (Archana & Senthil Kumar, 2023), linear regression (Mohamad Mohsin et al., 2024), as well as classification and regression-based approaches such as decision trees and random forests (Moraye et al., 2021). Nandani and Vidanapathirana (2024) demonstrated the potential of an artificial neural network model trained with backpropagation learning, where the Batch Gradient Descent (BGD) optimisation method yielded the most accurate paddy yield forecasts based on climatic variables. Other methods employed for predicting rice yields include statistical approaches (Guruprasad et al., 2019), machine learning (Ahmad et al., 2017; Alfred et al., 2021; Reddy et al., 2023), and remote sensing with satellite imagery analysis (Liu et al., 2022). Based on the literature, a rice prediction model needs to be simple and robust to integrate multiple climatic and agricultural variables in solving the prediction problem (Attri et al., 2023). One of the advantages of integrating remote sensing into a prediction model is that it provides valuable crop health and environmental data, enables large-scale monitoring of agricultural fields, assesses vegetation indices, and identifies stress factors such as drought or pest infestations. This integration of spatial and temporal data enhances predictive capabilities and complements traditional and machine learning-based models for yield estimation (Ali et al., 2022; You et al., 2017). Furthermore, Maya Gopal and Bhargavi (2019) and Nigam et al. (2019) highlight that adapting machine learning with different variables can promote a promising result.

Recent studies suggested that the combination of process-based and data-driven models could improve prediction accuracy and practical utility. Through a hybrid approach, it integrates the explainability feature of process-based approaches with the computational power of machine learning. For instance, Feng et al. (2020) integrated APSIM with regression modelling to dynamically track climate indices. Besides that, Shahhosseini et al. (2021) demonstrated enhanced predictive performance by coupling APSIM with machine learning techniques. These findings highlight the potential for hybrid approaches in advancing agricultural yield prediction.

### Climate and Rice Plantation

The main climatic variables are temperature, humidity, rainfall, and wind speed. These variables significantly impact rice yield because they govern physiological processes like photosynthesis, respiration, and grain filling (Peng et al., 2004). High temperatures during critical growth stages, particularly flowering, can reduce rice yield by causing sterility (Wassmann et al., 2009). Similarly, excessive rainfall or prolonged droughts can disrupt planting cycles, while high humidity affects plant transpiration rates and soil moisture retention (Hatfield & Prueger, 2015).

Climatic variables and rice yield exhibit a non-linear relationship, and they are characterised by thresholds beyond which yields decline sharply (Schlenker & Roberts, 2009). For example, moderate increases in temperature can boost yield within an optimal range but cause significant reductions once temperatures exceed a critical threshold (Challinor et al., 2007). These complexities underscore the need for models that account for non-linear interactions, like polynomial features, to improve predictive accuracy. Incorporating non-linear features into regression models enables the representation of curvilinear effects. This enhancement enhances the model’s ability to reflect real-world patterns accurately. Wassmann et al. (2009) demonstrated that quadratic terms for temperature improved predictions of yield variability under climate stress. Similarly, Lobell et al. (2008) emphasised the importance of considering interactions between temperature and humidity, as their combined effects have a significant impact on crop outcomes.

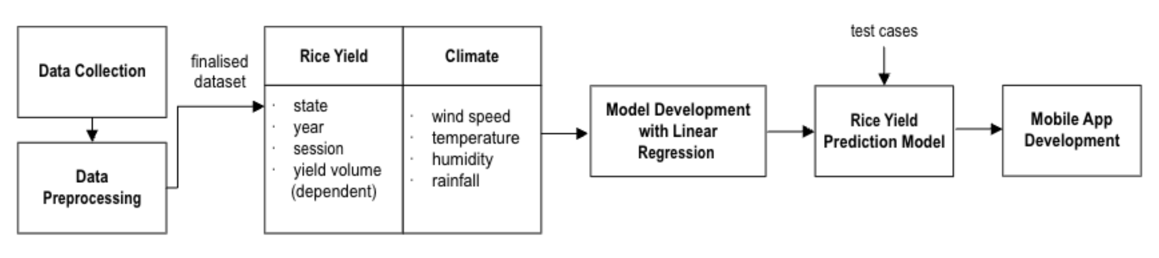
By integrating polynomial and interaction terms, rice yield prediction models can have a collective impact, more accurately reflecting the complex relationships between climatic variables and their combined effect on agricultural productivity. This approach bridges the gap in traditional linear methods and aligns with recent advancements in agricultural modelling.

## METHODOLOGY

This study adopted a data mining research methodology. It consists of five phases, which start with data collection, data pre-processing, model development, evaluation, and conclude with application development. Developing a mobile application to implement and test a proposed prediction model is essential for validating its practical effectiveness. While a model may perform well in simulations, a mobile application enables real-world testing, ensuring it can effectively handle both real-time and batch data. It also provides a structured way to measure key performance metrics, such as accuracy, precision, recall, and execution time, while identifying practical constraints, including computational efficiency and latency. Additionally, a mobile application enables usability testing, allowing stakeholders to interact with the model and assess its interpretability. Figure 1 depicts the methodology of this study.

**Figure 1**

*Research Methodology*



## Data Collection

The input dataset for this study was collected from two government agencies in Malaysia. Previous rice yield production information from 2010 to 2021 was obtained from the DOA. This dataset includes six variables that consist of yield (numerical), parcel area (numerical), planted area (numerical), state (categorical), year (categorical), and season (categorical). Yield data represent annual information on each state. The planted area indicates the actual rice growing area, while the parcel area denotes the rice field utilised for planting the rice. Meanwhile, the state information gives valuable insight into land utilisation for rice farming. The corresponding climate data set was sourced from MetMalaysia. The dataset comprises monthly records of wind speed, temperature, humidity, and rainfall for Malaysian states from 2010 to 2021. These variables are risk factors in rice plantations as they influence plant growth and development.

## Data Pre-processing

Both datasets underwent pre-processing to ensure their quality for multiple regression analysis. A seasonal imputation technique addressed missing data points in the climate dataset, where the missing values were filled with the mean of the corresponding month from the previous and succeeding years, as shown in Equation 1. In this equation, *m* represents the corresponding month.

$$Climate_m = \frac{Climate_{m-1} + Climate_{m+1}}{2} \tag{1}$$

Both datasets were transformed from monthly to yearly values to ensure a consistent time frame. A summation method was applied for the rice yield data to aggregate monthly values into a single value representing the annual yield. This approach captures the total yield output within a year, reflecting the biannual harvesting cycles in Malaysia. The model accounts for yield variations between in-season and off-season periods and different granaries by aggregating monthly data. The “season” and “state” variables are considered during modelling to provide a comprehensive view of annual yield performance.

Meanwhile, for climate data, the median value for each variable (rainfall, humidity, wind speed, temperature) was chosen to represent the entire year, as it is more robust to outliers than the average. The climate data was then merged with the rice yield data based on state and year, ensuring that each rice yield data point had corresponding climate data from the exact location and year to analyse the relationship between climate and rice yield effectively. The merging process involved aligning the datasets using common state and year fields. After merging both datasets, the total number of records ready for development is 312. The merged dataset is referred to as the rice yield dataset.

Having all features on a similar scale is recommended to ensure accurate and stable prediction results in regression modelling. Without normalisation, features with larger numerical ranges, such as temperature measured in degrees or rainfall in millimetres, can dominate the learning process and disproportionately influence the regression coefficients, potentially leading to biased or misleading predictions. To address this, the numerical features (climate variables) were normalised using the MinMaxScaler approach, which rescales all values to a uniform range between 0 and 1. It ensures that each feature contributes equally during model training, regardless of its original scale. Meanwhile, one-hot encoding was applied to the categorical variable “state.” Each state was transformed into a binary vector representation, allowing the model to treat each category independently. This approach avoids introducing unintended ordinal relationships and ensures that the model does not infer a false hierarchy among the states.

Polynomial features and interaction terms were added to the dataset to enhance the model’s ability to represent non-linear interactions. Specifically, squared terms for temperature and humidity were included to capture curvilinear effects—such as how moderate levels may enhance yield while extreme levels may reduce it. The additional features introduced are quadratic terms (Temperature$^2$, Humidity$^2$) and interaction terms (Temperature * Humidity). This transformation improves the model’s capacity to capture complex climatic effects. For example, the quadratic term, where moderate temperatures may boost yield, but extreme values reduce it. The new features were integrated into the rice yield dataset for model development.

To ensure robust model evaluation, the final step involved partitioning the experiment data into training and testing groups with varying proportions. The percentage of both groups is 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, and 80:20. The objective is to provide a comprehensive performance model evaluation across varying data availability scenarios, which can enhance dependability and robustness. It started with training on the designated training fold, followed by evaluation on the corresponding held-out testing fold for every ratio. Table 1 shows a sample of the pre-processed rice yield dataset.

**Table 1**

*The Pre-processed Rice Yield Dataset*

| Variables | Sample Data | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Median Rainfall | 0.4 | 0.45 | 0.46 | 0.44 | 0.42 |
| Median Humidity | 0.6 | 0.65 | 0.68 | 0.64 | 0.54 |
| Median Windspeed | 0.47 | 0.5 | 0.4 | 0.4 | 0.47 |
| Median Temperature | 0.8 | 0.8 | 0.81 | 0.85 | 0.86 |
| Crop Area (Ha) | 0.008205 | 0.008326 | 0.007743 | 0.008612 | 0.00889 |
| Parcel Area | 0.004308 | 0.002045 | 0.002045 | 0.002219 | 0.002219 |
| Outside Season | 0 | 0 | 0 | 0 | 0 |

(continued)

| Variables | Sample Data | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Main Season | 1 | 1 | 1 | 1 | 1 |
| Johor | 1 | 1 | 1 | 1 | 1 |
| Kedah | 0 | 0 | 0 | 0 | 0 |
| Kelantan | 0 | 0 | 0 | 0 | 0 |
| Melaka | 0 | 0 | 0 | 0 | 0 |
| Negeri Sembilan | 0 | 0 | 0 | 0 | 0 |
| Pahang | 0 | 0 | 0 | 0 | 0 |
| Perak | 0 | 0 | 0 | 0 | 0 |
| Perlis | 0 | 0 | 0 | 0 | 0 |
| Penang | 0 | 0 | 0 | 0 | 0 |
| Terengganu | 0 | 0 | 0 | 0 | 0 |
| Yield (Ton) | 5572 | 5560 | 5442 | 6031 | 6987 |

## Model Development

Building upon the pre-processed data, the next step involved model development. In this phase, multiple linear regression was employed as the baseline approach for predicting rice yield due to its simplicity, interpretability, and reliable performance compared to more complex models, such as decision trees and neural networks (Kaur et al., 2024; Xie et al., 2024). This model was a foundation for exploring enhancements to capture more intricate relationships between variables.

The prediction of rice yield, expressed in tons, was undertaken using a multiple linear regression model incorporating eight independent variables: rainfall, median, wind speed, temperature, crop area, pasar area, season type, and state. During the training, the multiple linear regression establishes a mathematical relationship to map the target variable (rice yield amount) with the influencing factors (climate and rice yield information). Polynomial features and interaction terms were incorporated to enhance the predictor's ability to capture complex climatic relationships. The regression model with these enhancements is described in Equation 2:

$$Y = b_1 X_1 + b_2 X_2 + b_3 X_1^2 + b_4 X_2^2 + b_5 (X_1 \times X_2) + b_6 X_6 + \ldots. + b_n X_n + a + e \tag{2}$$

where,
Y: Predicted rice yield
$X_1, X_2, X_3, \ldots., X_n$: Predictor variables (temperature, humidity, rainfall, wind speed)
$b_1$ to $b_n$: Regression coefficients for each input variable
$X_1^2, X_2^2, X_1 \times X_2$: Polynomial and interaction terms
a: Intercept term
e: Error term representing the discrepancy of rice prediction value

The inclusion of polynomial terms, such as $X_1^2$ (temperature squared) and $X_2^2$ (humidity squared), allows the model to account for non-linear effects where moderate levels may improve yield, but extreme values might reduce it. Interaction terms ($X_1 \times X_2$) capture the combined impact of temperature and humidity, which significantly affects rice growth and development. Polynomial terms were applied to temperature and humidity due to their well-documented non-linear effects on rice yield. Moderate temperature increases under optimal humidity levels can boost yield, while extremes reduce it (Schlenker & Roberts, 2009). In contrast, rainfall and wind speed generally exhibit more direct or linear relationships, making polynomial representation less critical in this context. Table 2 presents a sample of the transformed features for temperature and humidity and the interaction terms and polynomial terms of both features.

**Table 2**

*A Sample of the Transformed Temperature and Humidity Features, Including Interaction and Polynomial Terms*

| No | Median Temperature | Median Humidity | Temp_Humidity Interaction | Temperature Squared | Humidity Squared |
|---|---|---|---|---|---|
| 1 | 0.81 | 0.60 | 0.82 | 0.77 | 0.57 |
| 2 | 0.80 | 0.65 | 0.84 | 0.76 | 0.63 |
| 3 | 0.81 | 0.68 | 0.87 | 0.78 | 0.66 |
| 4 | 0.85 | 0.64 | 0.88 | 0.82 | 0.62 |
| 5 | 0.86 | 0.54 | 0.85 | 0.84 | 0.52 |
| 6 | 0.89 | 0.54 | 0.88 | 0.87 | 0.51 |
| 7 | 0.91 | 0.54 | 0.90 | 0.89 | 0.52 |
| 8 | 0.87 | 0.61 | 0.89 | 0.84 | 0.59 |
| 9 | 0.84 | 0.58 | 0.84 | 0.81 | 0.55 |
| 10 | 0.90 | 0.46 | 0.84 | 0.87 | 0.44 |

To ensure the model generalises well on unseen data, the best regression line for prediction is chosen from the train-test splitting sets during model development. Four error-measurement metrics assess the model's performance and include the following (Chicco et al., 2021):

i. Mean Absolute Error (MAE) captures the residual of the predictive error. It presents it as an average of absolute prediction error size, regardless of whether it is an overestimation (positive) or an underestimation (negative).
ii. Mean Squared Error (MSE) emphasises significant incorrect predictions by squaring the differences from MAE, ensuring that more significant errors substantially impact the overall statistic.
iii. Root Mean Squared Error (RMSE) is similar to MSE, but it transforms the score back to the original units (e.g., tons of rice) for a more straightforward interpretation.
iv. R-squared ($R^2$) represents the proportion of the variability in the rice yield feature described by the independent variables.

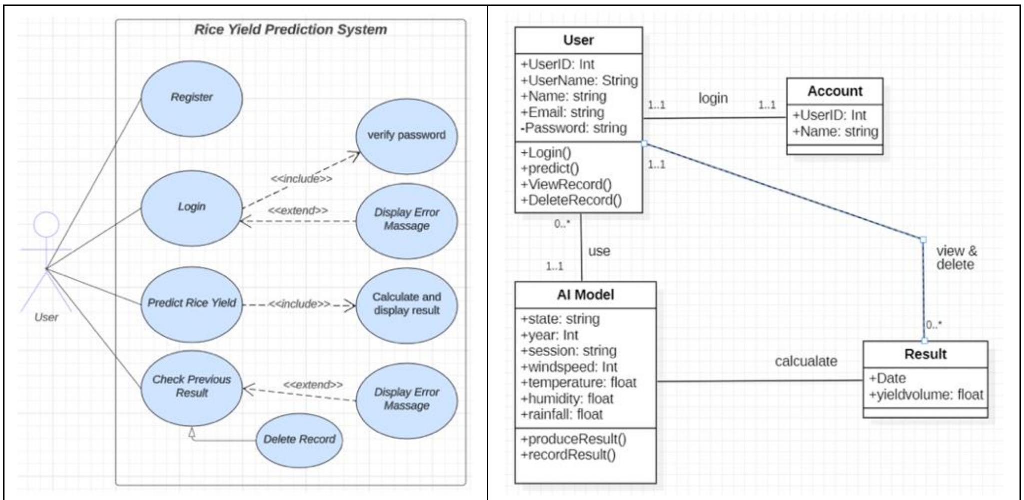
For MAE, MSE, and RMSE, the lower values indicate better agreement between predicted and actual yields. Meanwhile, the R2 is between 0 and 1, indicating a better fit as it approaches 1 (perfect).

### Mobile Application Development

The final phase involved developing a user-friendly mobile application to leverage the created rice yield prediction model. Without mobile application development, a predictive model remains a theoretical concept, making implementation crucial in proving its effectiveness. This section outlines the selected technologies and the overall mobile application design. Use case and database diagrams were employed to understand system functionalities and data management. Figure 2(a) is the use case diagram, which depicts the core functionalities from the user's perspective. This includes actors (e.g., stakeholders) interacting with the system by providing field data, initiating predictions, and viewing the generated yield estimates. A database diagram, shown in Figure 2(b), illustrates the data structure used within the system. This schema includes tables for field yield and climate data and the prediction results for future reference.

**Figure 2**

*MRYPA Design Models*



*(a) Use Case Diagram* *(b) Database Diagram for the MRYPA*

The user interface was built using Flutter, a popular mobile app framework, and the Dart programming language. The Model-View-Controller design pattern is employed in this mobile application to organise the code structure. To facilitate communication between the user interface and the AI model, a custom API built using the Flask framework and RESTful principles acts as an intermediary. Additionally, a database management system managed by phpMyAdmin is employed to store user data and past prediction results. Finally, the completed mobile application is deployed to a web hosting server for online accessibility. By effectively utilising these diagrams and a well-chosen technology stack, the development process ensured a clear understanding of system functionalities, data management, and overall mobile application design.

## EVALUATION AND RESULTS

This section presents the key findings of this study, encompassing both the developed prediction model and the mobile rice yield prediction application.

## Prediction Model Performance

This section presents the comparative performance of the baseline (standard multiple linear regression) and enhanced (polynomial and interaction terms) prediction models. To ensure the models generalise well to unseen data and avoid overfitting, the train-test splitting technique was employed during model development, as described in the data pre-processing section. The data was randomly divided into eight folds, with test sizes ranging from 10% to 80% in increments of 10%. Each fold underwent training-testing to evaluate model performance across different data subsets. **Table 3** compares the prediction errors for rice yield across the baseline and enhanced models, including MAE, MSE, and RMSE. To visualise the differences, **Figure 3** illustrates the MAE values for both models across different test sizes, while **Figure 4** compares the R$^2$ values for the baseline and enhanced models.

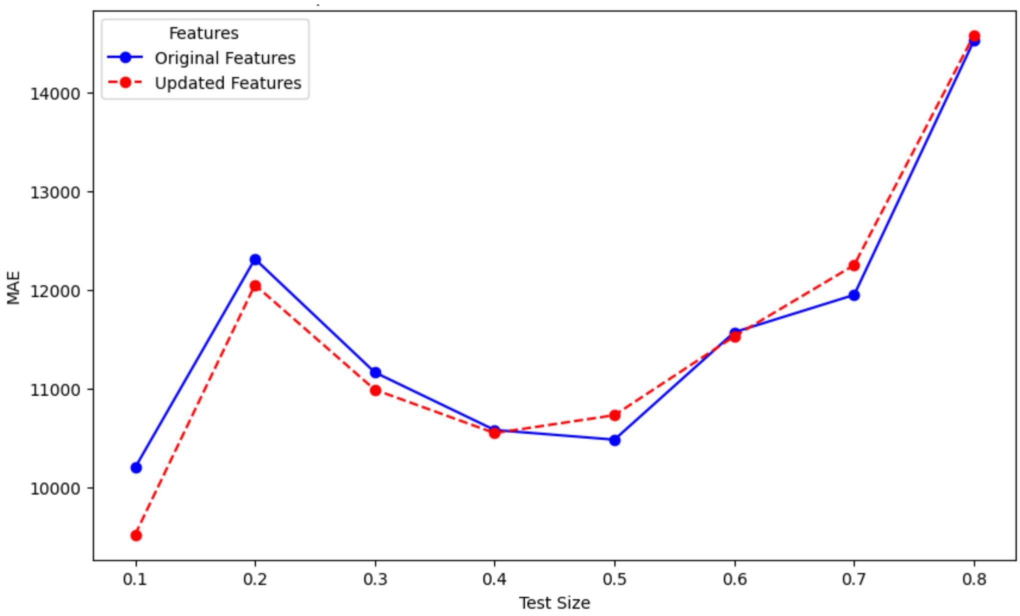
**Table 3**

*Error Metrics for Rice Yield Prediction in the Base Model*

| Test Size | Baseline (Standard Multiple Linear Regression) | | | Enhanced (Polynomial and Interaction Terms) | | |
|---|---|---|---|---|---|---|
| | MAE | MSE | RMSE | MAE | MSE | RMSE |
| 0.1 | 10,195.67 | 179,717,994.22 | 13,405.89 | 9,516.48 | 168,100,620.40 | 12,965.36 |
| 0.2 | 12,310.97 | 325,585,024.74 | 18,043.97 | 12,048.11 | 323,588,084.98 | 17,988.55 |
| 0.3 | 11,161.86 | 288,288,400.05 | 16,979.06 | 10,986.10 | 284,980,338.75 | 16,881.36 |
| 0.4 | 10,577.38 | 287,447,432.36 | 16,954.27 | 10,546.13 | 283,731,400.02 | 16,844.33 |
| 0.5 | 10,479.38 | 254,729,041.80 | 15,960.23 | 10,727.58 | 256,350,263.63 | 16,010.94 |
| 0.6 | 11,568.49 | 299,431,174.63 | 17,304.08 | 11,526.14 | 299,425,382.59 | 17,303.91 |
| 0.7 | 11,946.77 | 308,069,811.40 | 17,551.92 | 12,248.76 | 318,202,381.66 | 17,838.23 |
| 0.8 | 14,525.24 | 429,819,860.47 | 20,732.10 | 14,580.39 | 418,308,643.19 | 20,452.60 |

**Figure 3**

*MAE for Baseline (Original Feature) and the Enhanced Model (Updated Features)*



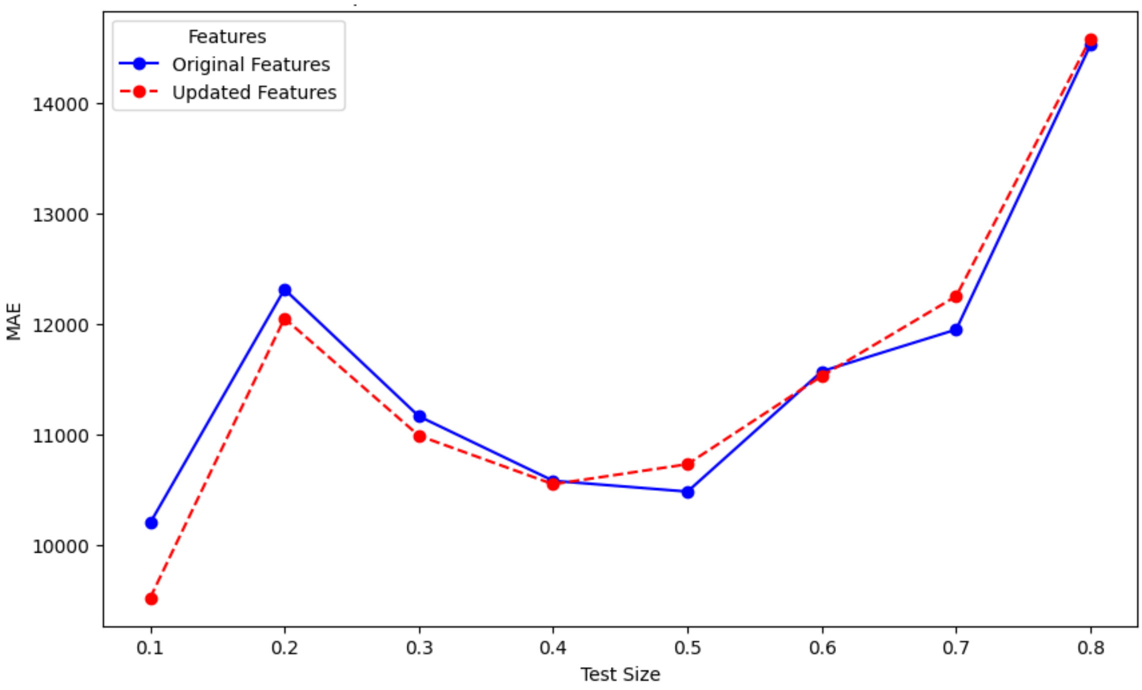
Both models exhibit similar performance trends across different test sizes, with MAE and RMSE generally increasing as the test size increases, reflecting the reduced amount of training data available for model development. As shown in Table 3 and Figure 3, the enhanced model, which incorporates polynomial and interaction terms, consistently demonstrates slight improvements in MAE and RMSE for smaller test sizes, indicating marginally better predictive performance when more training data is available.

In scenarios with smaller test sizes (10%-30%), the enhanced model outperforms the baseline model in all metrics. This suggests that incorporating polynomial and interaction terms helps capture non-linear relationships and interactions among variables, especially when sufficient training data is available. However, as test sizes increase beyond 50%, the performance gap between the models narrows, and in some instances, such as at 70% test size, the baseline model performs marginally better. This observation indicates that the added complexity of the enhanced model does not always translate into improved performance when training data is significantly reduced, highlighting the importance of balancing model complexity with data availability.

Beyond MAE and RMSE, the R² metric helps interpret the predictor's ability to describe the variability in the dependent variable. The R² with a high value (approaching 1) signifies a better fit, while a lower value (approaching 0) implies a weaker relationship. While MAE and RMSE measure prediction accuracy in terms of error magnitude, R² evaluates the predictor's ability to generalise relations exits in the data. These metrics provide a holistic view of model performance, as shown in Figure 4.

**Figure 4**

*$R^2$ for Baseline (Original Features) and the Enhanced Model (Updated Features)*



As shown in Figure 4, the enhanced model demonstrates marginally higher R² values across most test sizes, indicating a better overall fit for the data. This reinforces the earlier observation that the enhanced model can capture more complex relationships, particularly in scenarios with smaller test sizes. For smaller test sizes (10%-30%), the enhanced model consistently achieves higher $R^2$ values, confirming its effectiveness in leveraging polynomial and interaction terms to improve predictive accuracy. However, at larger test sizes (50% and above), the $R^2$ values for both models converge, indicating that the baseline model's simplicity is sufficient when more training data is available.

Combining $R^2$ with MAE and RMSE makes it evident that while the enhanced model offers advantages in smaller datasets by capturing non-linear relationships, its added complexity may not yield significant benefits when larger training datasets are utilised. The enhanced model, with a 50% train-test split, demonstrated marginally superior performance, exhibiting the smallest MAE, MSE, and RMSE values, along with a slightly higher R-squared value than the baseline model. The observation shows that the enhanced version has a slightly better ability to capture non-linear relationships and interactions among variables, particularly when a balanced dataset split is used. However, the performance gap between the enhanced and baseline models is modest, suggesting that the baseline model's simplicity suffices for comparable results when computational efficiency is a priority.

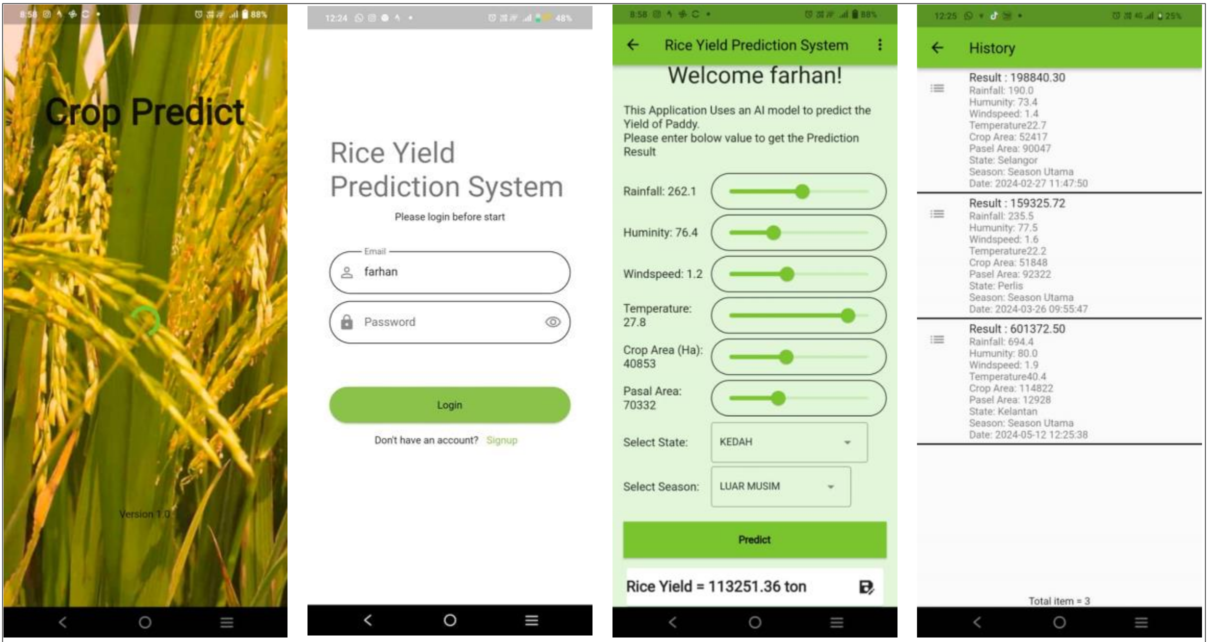
### Mobile Rice Yield Prediction Application

Building upon the promising results of the rice yield prediction model, a user-friendly mobile mobile application was developed to leverage its capabilities. This section details the design and functionalities of the MRYPA. Figures 5(a)- 5 (d) show the mobile application's interface. The mobile application prioritises user security by implementing a login interface. This initial screen allows authorised users and agricultural stakeholders to access the system's functionalities (see Figure 5(b)). Users can interact with the mobile application through a user-friendly interface upon successful login. Users can navigate the system through data input, data submission, receiving results, and recording predictions (see Figure 5(c). The data input module provides a streamlined platform for users to enter critical information about their rice fields. This includes yield properties such as crop area, parcel area, state, and season, along with climate conditions like rainfall, humidity, wind speed, and temperature. Once the necessary data is entered, users can proceed with the data submission process, which involves initiating the prediction model via a designated action within the interface.

Upon receiving the input, the mobile application predicts the rice yield (in tons) using the trained prediction model built into the mobile application. The results are, offering users valuable insights into potential harvest outcomes. Additionally, the mobile application includes a data storage feature, allowing users to securely store their field data and past prediction results in a database. This functionality facilitates future reference, enables historical data analysis, and supports comparisons with actual yields (see Figure 5(d)). The mobile rice yield prediction application empowers farmers and agricultural stakeholders with valuable insights by offering a secure login process, a user-friendly interface, seamless data submission, and a precise prediction display.

**Figure 5**

*MRYPA Interface's*



farhan

Welcome farhan!

This Application Uses an AI model to predict the Yield of Paddy.
Please enter bolow value to get the Prediction Result

Rainfall: 262.1
Huminity: 76.4
Windspeed: 1.2
Temperature: 27.8
Crop Area (Ha): 40853
Pasal Area: 70332
Select State: KEDAH
Select Season: LUAR MUSIM

Rice Yield = 113251.36 ton

Result : 198840.30
Rainfall: 190.0
Humunity: 73.4
Windspeed: 1.4
Temperature22.7
Crop Area: 52417
Pasel Area: 90047
State: Selangor
Season: Season Utama
Date: 2024-02-27 11:47:50

Result : 159325.72
Rainfall: 235.5
Humunity: 77.5
Windspeed: 1.6
Temperature22.2
Crop Area: 51848
Pasel Area: 92322
State: Perlis
Season: Season Utama
Date: 2024-03-26 09:55:47

Result : 601372.50
Rainfall: 694.4
Humunity: 80.0
Windspeed: 1.9
Temperature40.4
Crop Area: 114822
Pasel Area: 12928
State: Kelantan
Season: Season Utama
Date: 2024-05-12 12:25:38

Total item = 3

*(a) Welcome Screen* *(b) Login* *(c) Prediction Function* *(d) History Information*

## Implications of the Result

The results from this study demonstrate that incorporating polynomial and interaction terms into a traditional linear regression model can lead to improved predictive performance, especially in scenarios with limited training data. This has practical implications for small-scale or region-specific agricultural systems, where large datasets may not always be available. The enhanced model's ability to capture non-linear relationships is particularly beneficial in modelling the complex interactions between climatic variables and crop yield. However, the relatively small performance gap observed at larger test sizes also highlights the strength of simpler models in terms of computational efficiency and ease of deployment. These findings suggest that model selection should be context-dependent, balancing the trade-off between accuracy and resource constraints. For practitioners, this means that enhanced models are suitable where precision is critical and data is limited, while baseline models may suffice in large-scale or real-time applications.

## CONCLUSION

This study proposed and evaluated an enhanced rice yield prediction model that incorporates polynomial and interaction terms to better capture non-linear climatic effects. By extending the baseline multiple linear regression model, the enhanced approach aimed to address limitations in traditional linear predictions, particularly under complex environmental conditions. The MRYPA integrates a comprehensive dataset that includes crop properties (e.g., area, parcel area, state, season) and critical climatic factors (e.g., rainfall, humidity, wind speed, temperature). The enhanced model demonstrated marginal improvements in performance, particularly at smaller test sizes (10%-30%), where it effectively captured non-linear relationships and interactions among variables. With a 50% test size split, both models achieved strong performance, with the enhanced model yielding slightly better metrics, including an R-squared value of 0.9843.

However, as test sizes increased, the baseline model's simplicity proved equally sufficient, highlighting the importance of balancing model complexity with data availability. Given the significant role these factors play in influencing rice yield, incorporating climate data was essential for enhancing model accuracy. While the improved model showed promise in capturing additional dynamics, the modest performance gains suggest a need to explore other approaches to address model limitations. These findings imply that enhanced models are more suitable for scenarios with limited data, whereas baseline models offer practical advantages in larger-scale or resource-constrained settings. Future work will involve integrating additional data sources such as soil properties, pest data, and socioeconomic factors to further refine the prediction capability. Moreover, considering advanced machine learning techniques such as ensemble and hybrid models may help improve predictive accuracy. Risk factors such as pandemics (e.g., COVID-19) will also be incorporated alongside climate data to enhance the application's robustness and applicability. This approach aims to support stakeholders with a more resilient, accurate, and user-friendly decision-making tool, supporting sustainable farming and resource optimisation to address dynamic challenges.

## ACKNOWLEDGMENT

The authors of this article gratefully acknowledge the support provided by the Ministry of Higher Education (MoHE) through the Fundamental Research Grant Scheme (Ref: FRGS/1/2021/SS02/UUM/02/1 (S/O Code: 20107)). However, the views expressed in this article are those of the authors alone and do not necessarily reflect the official position of the MoHE, Malaysia.

## REFERENCES

Abidin, A. Z. Z., & Dardak, R. A. (2023). Sociological issues and challenges of rice production in Malaysia. *FFTC Agricultural Policy Platform.* https://ap.fftc.org.tw/article/3473

Ahmad, A. A., Shitan, M., & Yusof, F. (2017). Forecast of annual paddy production in MADA region using ARIMA (0,2,2) model. *Economic and Technology Management Review*, *12*, 11–17.

Alam, M. M., Siwar, C., Jaafar, A. H., Talib, B., & Salleh, K. O. (2013). Climate change adaptability of farmers: Malaysian case study. *International Journal of Plant, Animal and Environmental Sciences, 3*(3), 130–135.

Alfred, R., Obit, J. H., Chin, C. P.-Y., Haviluddin, H., & Lim, Y. (2021). Towards paddy rice smart farming: A review on big data, machine learning, and rice production tasks. *IEEE Access*, *9*, 50358–50380. https://doi.org/10.1109/ACCESS.2021.3069449

Ali, A. M., Abouelghar, M., Belal, A. A., Saleh, N., Yones, M., Selim, A. I., Amin, M. E. S., Elwesemy, A., Kucher, D. E., Maginan, S., & Savin, I. (2022). Crop yield prediction using multi sensors remote sensing (Review Article). *The Egyptian Journal of Remote Sensing and Space Science*, *25*(3), 711–716. https://doi.org/10.1016/j.ejrs.2022.04.006

Archana, S., & Senthil Kumar, P. (2023). A survey on deep learning-based crop yield prediction. *Nature Environment and Pollution Technology*, *22*(2), 579–592. https://doi.org/10.46488/NEPT.2023.v22i02.004

Attri, I., Awasthi, L. K., Sharma, T. P., & Rathee, P. (2023). A review of deep learning techniques used in agriculture. *Ecological Informatics*, *77*, 102217. https://doi.org/10.1016/j.ecoinf.2023.102217

Baharudin, S. 'Aisyah, & Waked, H. N. (2021). Machinery and technical efficiencies in selected paddy areas in Malaysia. *Pertanika Journal of Social Sciences and Humanities*, *29*(4), 2225–2242. https://doi.org/10.47836/pjssh.29.4.07

Balboa, G. R., Archontoulis, S. V., Salvagiotti, F., Garcia, F. O., Stewart, W. M., Francisco, E., Vara Prasad, P. V. & Ciampitti, I. A. (2019). A systems-level yield gap assessment of maise-soybean rotation under high-and low-management inputs in the Western US Corn Belt using APSIM. *Agricultural Systems*, *174*, 145-154. https://doi.org/10.1016/j.agsy.2019.04.008

Challinor, A. J., Wheeler, T. R., Craufurd, P. Q., Ferro, C. A. T., & Stephenson, D. B. (2007). Adaptation of crops to climate change through genotypic responses to mean and extreme temperatures. *Agriculture, Ecosystems & Environment*, *119*(1–2), 190–204. https://doi.org/10.1016/j.agee.2006.07.009

Chang, J., Ma, L., Hoogenboom, G., & Ahuja, L. R. (2023). Bridging process-based and machine-learning crop models to improve prediction accuracy and adaptivity. *Communications Biology*, *6*(1), 1–9. https://doi.org/10.1038/s42003-023-04833-y

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Corbeels, M., Chirat, G., Messad, S. & Thierfelder, C. (2016). Performance and sensitivity of the DSSAT crop growth model in simulating maise yield under conservation agriculture. *European Journal of Agronomy*, *76*, 41–53. https://doi.org/10.1016/j.eja.2016.02.001

Dorairaj, D., & Govender, N. T. (2023). Rice and paddy industry in Malaysia: Governance and policies, research trends, technology adoption and resilience. *Frontiers in Sustainable Food Systems*, *7*. https://doi.org/10.3389/fsufs.2023.1093605

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

Firdaus, R. B. R., Leong Tan, M., Rahmat, S. R., & Senevi Gunaratne, M. (2020). Paddy, rice and food security in Malaysia: A review of climate change impacts. *Cogent Social Sciences*, *6*(1). https://doi.org/10.1080/23311886.2020.1818373

Guruprasad, R. B., Saurav, K., & Randhawa, S. (2019). Machine learning methodologies for paddy yield estimation in India: A case study. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 7254–7257. https://doi.org/10.1109/IGARSS.2019.8900339

Hatfield, J. L., & Prueger, J. H. (2015). Temperature extremes: Effect on plant growth and development. *Weather and Climate Extremes*, *1*0, 4-10. http://dx.doi.org/10.1016/j.wace.2015.08.001

Herman, T., Murchie, E. H., & Warsi, A. A. (2015). Rice production and climate change: A case study of Malaysian rice. *Pertanika Journal of Tropical Agricultural Science*, *38*(3), 321–328.

Jabin, N., Amin, A. Q. Al, Kari, F., & Alam, G. M. (2015). Climate change adaptation provisions for the agricultural sector in Malaysia. *International Journal of Global Warming*, *7*(3), 336. https://doi.org/10.1504/IJGW.2015.069366

Nandani, E. J. K. P., & Vidanapathirana, T. T. S. (2024). Forecasting paddy yield in Sri Lanka using back-propagation learning in artificial neural network model. *Journal of the University of Ruhuna*, *12*(2). 110-120. https://doi.org/10.4038/jur.v12i2.8032

Kaur, H., Agrawal, R., & Kumari, M. (2024). Rice yield prediction through integration of biophysical parameters and machine learning models. *Scientific Reports*, *14*(1), Article 72624. https://doi.org/10.1038/s41598-024-72624-4

Lobell, D. B., Burke, M. B., Tebaldi, C., Mastrandrea, M. D., Falcon, W. P., & Naylor, R. L. (2008). Prioritising climate change adaptation needs for food security in 2030. *Science, 319*(5863), 607-610.

Liu, Y., Wang, S., Chen, J., Chen, B., Wang, X., Hao, D., & Sun, L. (2022). Rice yield prediction and model interpretation based on satellite and climatic indicators using a transformer method. *Remote Sensing*, *14*(19), 5045. https://doi.org/10.3390/rs14195045

Makhtar, S., Abidin, I. S. Z., & Islam, R. (2022). The impact of rice productivity in Malaysia: Econometric analysis. *International Journal of Business and Economy (IJBEC)*, *4*(2), 21–32.

Masud, M. M., Rahman, Md. S., Al-Amin, A. Q., Kari, F., & Filho, W. L. (2014). Impact of climate change: An empirical investigation of Malaysian rice production. *Mitigation and Adaptation Strategies for Global Change*, *19*(4), 431–444. https://doi.org/10.1007/s11027-012-9441-z

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

Mohd Ali, N. I., Ibrahim, N. I., Aiyub, K., Kasavan, S., Kuok Choy, L., & Siron, R. (2021). Challenges in implementation of MyGAP among paddy farmers. *Malaysian Journal of Society and Space, 17*(4). https://doi.org/10.17576/geo-2021-1704-12

Mohamad Mohsin, M. F., Umana, M. K., Hassan, M. G., Sharif, K. I. M., Ismail, M. A., Salleh, K., Mohd Zahari, S., Sarmani, M. A. & Gordon, N. (2024). Improving rice yield prediction accuracy using regression models with climate data. In *International Conference on Computing and Informatics* (pp. 258-267). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-9592-9_20

Moraye, K., Pavate, A., Nikam, S., & Thakkar, S. (2021). Crop yield prediction using random forest algorithm for major cities in Maharashtra state. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST), 9*(2), 2347-5552.

Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019). Crop yield prediction using machine learning algorithms. *2019 Fifth International Conference on Image Information Processing (ICIIP)*, 125–130. https://doi.org/10.1109/ICIIP47207.2019.8985951

Omar, S. C., Shaharudin, A., & Tumin, S. A. (2019). *The status of the paddy and rice industry in Malaysia*. Khazanah Research Institute. https://www.krinstitute.org/assets/contentMS/img/template/editor/20190409_RiceReport_Full%20Report_Final.pdf

Peng, S., Huang, J., Sheehy, J. E., Laza, R. C., Visperas, R. M., Zhong, X., Centeno, G.S., Khush, G.S. & Cassman, K. G. (2004). Rice yields decline with higher night temperature from global warming. *Proceedings of the National Academy of Sciences, 101*(27), 9971-9975. https://doi.org/10.1073/pnas.0403720101

Reddy, D. S., Babu, K. S., & Basha, S. M. (2024). Weather based paddy yield prediction using machine learning regression algorithms. *Journal of Agricultural Mechanics*, *26*(3). https://doi.org/10.54386/jam.v26i3.2598

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proceedings of the National Academy of sciences*, *106*(37), 15594-15598. https://doi.org/10.1073/pnas.0906865106

Siwar, C., Diana Mohd Idris, N., Yasar, M., & Morshed, G. (2014). Issues and challenges facing rice production and food security in the granary areas in the East Coast Economic Region (ECER), Malaysia. *Research Journal of Applied Sciences, Engineering and Technology*, *7*(4), 711–722. https://doi.org/10.19026/rjaset.7.307

Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific Reports*, *11*(1), 1–15. https://doi.org/10.1038/s41598-020-80820-1

Tan, B. T., Fam, P. S., Firdaus, R. B. R., Tan, M. L., & Gunaratne, M. S. (2021). Impact of climate change on rice yield in Malaysia: A panel data analysis. *Agriculture*, *11*(6), 569. https://doi.org/10.3390/agriculture11060569

Wassmann, R., Jagadish, S. V. K., Sumfleth, K., Pathak, H., Howell, G., Ismail, A., Serraj, R., Redona, E., Singh, R. K. & Heuer, S. (2009). Regional vulnerability of climate change impacts on Asian rice production and scope for adaptation. *Advances in Agronomy*, *102*, 91-133. https://doi.org/10.1016/S0065-2113(09)01003-7

Xie, C., Zhang, M., & Li, H. (2024). Hybrid deep neural networks with multi-tasking for rice yield prediction using remote sensing data. *Agriculture*, *14*(4), Article 513. https://doi.org/10.3390/agriculture14040513

You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). 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). https://doi.org/10.1609/aaai.v31i1.11172