MULTIDISCIPLINARY APPLIED RESEARCH AND INNOVATION
e-ISSN: 2773-4773

MARI

Vol. 6 No. 2 (2025) 1-17
https://publisher.uthm.edu.my/periodicals/index.php/mari

Smart Agriculture Economics and Engineering: Unveiling the Innovation Behind AI-Enhanced Rice Farming

Zun Liang Chuan[1]*, Tham Ren Sheng[2], Tan Chek Cheng[1], Abraham Lim Bing Sern[1], David Lau King Luen[1], Chong Yeh Sai[3]

[1] *Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuh Persiaran Tun Khalil Yaakob, 26300 Gambang, Pahang, MALAYSIA*
[2] *Faculty of Industrial Management, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuh Persiaran Tun Khalil Yaakob, 26300 Gambang, Pahang, MALAYSIA*
[3] *Ever AI Holdings Sdn Bhd, 12, Jalan Anggerik Aranda 31/170C, 40460 Kota Kemuning, Shah Alam, Selangor, MALAYSIA*

*Corresponding Author: chuanzl@umpsa.edu.my
DOI: https://doi.org/10.30880/mari.2025.06.02.001

Article Info

Received: 31 August 2024
Accepted: 31 December 2024
Available online: 20 February 2025

Keywords

Food security, Rice production prediction, AI-based predictive algorithm, Southeast Asia, modified Taguchi-based VIKOR, multi-criteria decision-making

Abstract

Food security challenges in Southeast Asia, across all income brackets, have been growing, according to the Food and Agriculture Organization (FAO) of the United Nations. This article introduced innovative Artificial Intelligence-based (AI-based) predictive algorithms for short-term rice production, utilizing the Cross Industry Standard Process for Data Mining (CRISP-DM) data science framework. The predictive algorithms integrated features addressing three food security dimensions: availability, accessibility, and stability, and identified key determinants in three clusters: atmospheric, socioeconomic, and farming practices. By employing the proposed innovative modified stacked Multiple Linear Regression-Support Vector Regression-based (MLR-SVR-based) algorithms, and ranking them utilizing the modified Taguchi-based VIseKriterijumska Optimizacija I Kompromisno Resenje (Taguchi-based VIKOR) multi-criteria decision-making algorithm, the analysis demonstrated high predictive accuracy even with limited data. The proposed AI-based predictive algorithm was utilized to forecast 5-year future rice production for Southeast Asia nations, yielding generally accurate results except for Cambodia (KHM). This research has significant implications for agriculture, food production, data analytics, and policymaking, potentially enhancing efficiency and innovation in agricultural operations.

1. Introduction

Food security, as defined by the United Nations Committee on World Food Security, entails ensuring that all individuals have consistent physical, social, and economic access to sufficient, safe, and nutritious food that meets their preferences and dietary needs for an active and healthy life. This concept encompasses four key dimensions: availability, accessibility, utilization, and stability [1]. Availability ensures the consistent presence and quality of food from various sources. Accessibility involves having the income or resources to obtain appropriate food, while utilization focuses on proper food utilization and storage, considering nutritional and health factors. Stability ensures that food remains consistently available without disruption from emergencies or shortages.

In Asia and parts of the Pacific, agriculture, especially rice production, is crucial, providing approximately 90% of staple foods [2]. However, challenges such as climate change, exponential population growth, global food inflation, technological advancements, and national and international social-environmental stressors have increased food security uncertainty, impacting Sustainable Development Goals (SDGs) such as No Poverty (SDG1), Zero Hunger (SDG2), and Good Health and Well-Being (SDG3). Rice cultivation is also linked to Decent Work and Economic Growth (SDG8). However, the Food and Agriculture Organization (FAO) of the United Nations [3] reports a rise in the 3-year average prevalence of severe and moderate food insecurity (in total population) in Southeast Asia, particularly in low-middle-income and upper-middle-income nations.

To foster economic growth and sustainability in Southeast Asia, various studies have investigated predictive algorithms in fields such as econometrics, mathematical multivariable regression, traditional statistical multivariable and multivariate regression, and Artificial Intelligence-based (AI-based) predictive algorithms. These studies primarily focus on the availability and accessibility dimensions of food security. For instance, Tan et al. [4] utilized a fixed-effect panel regression econometrics algorithm to investigate the association between rice production and atmospheric clustered determinants in Malaysia (MYS). Their results showed a strong fit for the main season. In contrast, Chuan et al. [5] utilized a random-effect panel regression econometrics algorithm to study the association between agricultural production of C3 plants and various atmospheric and non-atmospheric clustered determinants in MYS, finding a limited fit due to a low coefficient of determination.

The utilization of fixed-effect and random-effect panel regression econometrics algorithms is widespread in recent Southeast Asia studies on rice production, considering atmospheric and socioeconomics clustered determinants. These studies cover nations such as Cambodia (KHM), Laos (LAO), Myanmar (MMR), Philippines (PHL), Vietnam (VNM), Indonesia (IDN), MYS, Thailand (THA), and Brunei (BRN) [6], THA [7], IDN [8], and PHL [9]. However, this paper does not explore their applicability due to the reliance on cross-sectional data and several statistical assumptions. Additionally, these studies do not simultaneously account for clustered determinants, including atmospheric, socioeconomic, and farming practices in econometrics algorithm development.

In previous studies on Southeast Asia, researchers have utilized various mathematical and statistical predictive algorithms to model and forecast rice and paddy production. These predictive algorithms include the general circular model-based (GCM-based) [10], Multiple Linear Regression (MLR) [11]-[15], Multivariate Normal Distribution (MND) [14], and Multivariate Copula-Based (MCB) algorithms [14]. Koide et al. [10] evaluated the GCM-based mathematical multivariable algorithm for predicting rice production in the PHL, considering climatology-related determinants at various levels. While GCM-based algorithms effectively capture trends, their deterministic nature presents challenges in forecasting due to inherent uncertainties. To overcome these challenges, researchers have turned to probabilistic predictive algorithms, which offer more reliable predictions.

Bashir and Yuliana [11] proposed utilizing a probabilistic multivariable MLR algorithm to regress the linearized associations of socioeconomic and farming practices clustered determinants toward rice production and consumption in IDN. However, this study's assessment of multicollinearity assumptions utilizing a correlation matrix might not fully capture the severity of multicollinearity, potentially affecting the reliability of the predictive algorithms. Similarly, Win et al. [12] investigated the linear association between socioeconomic and farming practices clustered determinants and the hybrid rice production among farmers in MMR, utilizing a probabilistic multivariable MLR algorithm. This study raised concerns as it did not provide diagnostic checking for the MLR algorithm.

Idalisa et al. [13] also utilized a probabilistic multivariable MLR algorithm to regress the linear association between socioeconomic and farming practices clustered determinants, and rice production in MYS. They aimed to improve the parameter estimation utilizing the Conjugate Gradient (CG) method compared to the classical Ordinary Least Squares (OLS) method. However, the study's conclusion might be suboptimal due to contradictory performance measurements between CG and OLS methods, and the lack of diagnostic checking. Recently, Aprizkiyandari and Palupi [15] conducted a comparative analysis of econometrics and traditional statistical multivariable regression predictive algorithms. They compared the fixed-effect panel regression and probabilistic multivariable MLR algorithms, but this study also did not include diagnostic checking, raising questions about the reliability and validity of the developed algorithm.

In addition to probabilistic multivariable MLR algorithms, probabilistic multivariate predictive algorithms such as the MCB algorithm have been explored in literature for modeling and forecasting paddy production in Southeast Asia. Roslan et al. [14] applied the MCB algorithm to five Southeast Asia nations: MMR, VNM, IDN, MYS, and THA. They evaluated the predictive algorithm's performance utilizing two families of multivariate copulas: elliptical copula (normal and t), and Archimedean copula (Joe, Clayton, and Gumbel) families, and compared it with MLR and MND algorithms. Their analysis demonstrated that the MCB algorithm generally outperformed both MLR and MND algorithms. However, a universally effective MCB algorithm for all five nations was not identified. Despite its potential, the MCB algorithm was not pursued further in this study due to the preliminary results showed that the MLR algorithm did not violate the multicollinearity assumption, indicating a low correlation among determinants.

AI-based predictive algorithms have gained attention for their potential to enhance the accuracy and robustness of rice production forecasts in Southeast Asia. Researchers have explored these predictive algorithms through feature engineering, algorithm refinement, and parameter optimization since the teen 21st century. Saithanu et al. [16] introduced an innovative AI-driven approach to predict rice production in THA utilizing atmospheric and rice-type dummy determinants within a machine learning framework, specifically applying the MLR algorithm. A key limitation of their study was the sole reliance on performance metrics such as Root Mean Square Error (RMSE), and adherence to OLS assumptions.

David [17] compared various AI-based predictive algorithms, including Artificial Neural Network (ANN), MLR, and Random Forest (RF) algorithms, for predicting rice production in the PHL. The RF algorithm outperformed both ANN and MLR algorithms, though the study was limited by a lack of feature selection, impacting parsimony, reliability, and cost, as well as the lack of diagnostic checking for the MLR algorithm. Consequently, the RF algorithm was not considered further in this paper due to its limited practical interpretability, particularly concerning statistically significant determinants.

Chuan et al. [18] compared the effectiveness of the MLR and Multiple Nonlinear Regression (MNLR) algorithms, along with various hybrid wrapper-filter feature selection methods, for modeling and forecasting rice production in MYS. They incorporated atmospheric, socioeconomic, and farming practices clustered determinants, splitting the dataset into training and test sets utilizing the Pareto principle. Their analysis found the MLR algorithm to be more effective than the MNLR algorithm. Notably, AI Neural Network-based (NN-based) algorithms such as Gated Recurrent Unit-based (GRU-based) and Long Short-Term Memory-based (LSTM-based) architectures, were not considered in this study. Previous research [19]-[23] had primarily focused on prediction evaluation rather than forecasting, and faced limitations such as sample size constraints and computational complexities. These complexities include determining the optimal number of epochs, hidden layers, and neurons for GRU-based and LSTM-based architectures, which impacted their practical applicability for forecasting future paddy and rice production.

Given the effectiveness demonstrated by Chuan et al. [18], the principal objective of this study is to propose an innovative modified stacked ensemble multivariable AI-based predictive algorithm for predicting rice production across six low-middle-income and three upper-middle-income nations in Southeast Asia. This AI-based predictive algorithm integrates classical statistical regression and semi-parametric machine learning regression algorithms, specifically the MLR and Support Vector Regression-based (MLR-SVR-based) algorithms, utilizing the Cross Industry Standard Process for Data Mining (CRISP-DM) data science framework. The low-middle-income nations considered are KHM, LAO, MMR, PHL, Timor-Leste (TLS), and VNM. The upper-middle-income nations included IDN, MYS, and THA.

This study makes significant contributions to both theoretical and practical domains. Theoretically, the proposed modified stacked ensemble AI-based predictive algorithm advances existing literature by enhancing predictive accuracy and interpretability. It effectively addresses three dimensions of food security: availability, accessibility, and stability, surpassing previous approaches. Practically, the predictive insights from this study support better health and well-being by ensuring access to sufficient and nutritious food. They also aid increase small farmers' income through improved rice production and promote agriculture entrepreneurs via sustainable practices. The research outcome facilitates informed decision-making and effective policy development, contributing to economic growth and sustainability in Southeast Asia.

2. Research Methodology

The CRISP-DM data science framework, well-regarded in both academia and industry, has evolved from its initial focus on optimizing business value to also supporting academic research with diverse applications, including commercialization pathways. This article explores its academic application within agriculture economics and engineering, building on its successful utilization in various fields such as computer science [24], energy economics [25], education economics [26]-[27], finance [28], and healthcare [29]. The CRISP-DM data science framework consists of six key phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase's statistical methodologies are tailored to the business and data mining objectives, the nature of the data, and the analysis results. The following sections detailed research methodologies specific to each phase within the context of this study.

2.1 Business Understanding

The business understanding phase aims to define both the business and data mining objectives, including research requirements, costs, risks, and management strategies. As outlined in Section 1, the business objective of this article is to provide predictive insights that optimize health-related products or services, enhance market competitiveness, improve consumer well-being, increase income opportunities for small farmers, support agricultural entrepreneurship, facilitate informed policy development, and boost agricultural productivity for economic growth and sustainability. Correspondingly, the data mining objective is to develop an innovative modified stacked ensemble multivariable AI-based predictive algorithm for predicting rice production across six low-middle-income and three upper-middle-income nations in Southeast Asia. To manage costs and risks, this study utilizes an open-source dataset and R statistical software, with minimal finance resources. To avoid suboptimal decisions, the modified Taguchi-based multi-criteria decision-making algorithm is utilized to rank the proposed and benchmark predictive algorithms due to conflicting Goodness-of-Fits (GoFs) measures. Figure 1 illustrates the management strategies based on the CRISP-DM data science framework to effectively achieve these objectives.

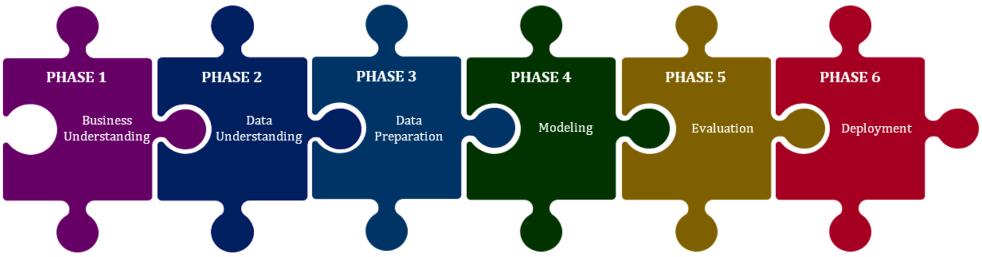


Fig. 1 *Unveiling management insights: CRISP-DM data science framework in focus*

2.2 Data Understanding

Understanding the data is foundational to comprehending the business context. This section provided insights into data collection, correlogram, correlation analysis, and exploratory data analysis (EDA). The dataset was sourced from reputable databases such as the Climate Change Knowledge Portal (CCKP) [30]-[31] and Our World in Data (OWID) [32]-[33], as documented in publications by leading publishers such as Elsevier. The dataset is presented in a Comma-Separated Value (CSV) file format spanning from 1961 to 2019. It encompassed 14 continuous determinants related to rice production as the endogenous variable ($\boldsymbol{\theta}_1$). These 14 determinants are classified into three principal categories: atmospheric ($\boldsymbol{\theta}_2$: annual mean maximum temperature; $\boldsymbol{\theta}_3$: average annual precipitation; $\boldsymbol{\theta}_4$: CO2 emission (per capita)), socioeconomic ($\boldsymbol{\theta}_5$: crude oil price (US$ per m^3); $\boldsymbol{\theta}_6$: domestic supply (tones); $\boldsymbol{\theta}_7$: food supply (kcal per capita per day); $\boldsymbol{\theta}_8$: import (tones); $\boldsymbol{\theta}_9$: inflation consumer prices (annual %); $\boldsymbol{\theta}_{10}$: labor index; $\boldsymbol{\theta}_{11}$: population; $\boldsymbol{\theta}_{12}$: urbanization rate (%)), and farming practices ($\boldsymbol{\theta}_{13}$: agriculture land index; $\boldsymbol{\theta}_{14}$: land use (hectare); $\boldsymbol{\theta}_{15}$: machinery per agriculture land). This study investigated the associations among these variables utilizing the correlogram and Pearson correlation coefficient, which revealed both direction and strength. This study highlighted that the high correlation among the determinants is not excluded from the data preparation and modeling phases. This is due to merely utilizing the correlation analysis in investigating the multicollinearity is insufficient. Additionally, hidden characteristics are explored through descriptive statistics, focusing on the first four L-moments (L-Mean (τ_1), L-Coefficient of Variation (τ_2), L-Skewness (τ_3), and L-Kurtosis (τ_4). L-moments are preferred over classical statistical moments due to their robustness against outliers and suitability for small sample sizes [25]. Mathematically, the first four sample L-moments [34] can be expressed as equations (1)-(4).

$$\tau_1 = \binom{T}{1}^{-1} \sum_{t=1}^{T} (\theta_i)_{(t:T)} \tag{1}$$

$$\tau_2 = \frac{1}{2}\binom{T}{2}^{-1} \sum_{t=1}^{T} \left\{ \binom{t-1}{1} - \binom{T-t}{1} \right\} (\theta_i)_{(t:T)} \tag{2}$$

$$\tau_3 = \frac{1}{3}\binom{T}{3}^{-1} \sum_{t=1}^{T} \left\{ \binom{t-1}{2} - 2\binom{t-1}{1}\binom{T-t}{1} + \binom{T-t}{2} \right\} (\theta_i)_{(t:T)} \tag{3}$$

$$\tau_4 = \frac{1}{4}\binom{T}{4}^{-1} \sum_{t=1}^{T} \left\{ \binom{t-1}{3} - 3\binom{t-1}{2}\binom{T-t}{1} + 3\binom{t-1}{1}\binom{T-t}{2} - \binom{T-t}{3} \right\} (\theta_i)_{(t:T)} \tag{4}$$

where $\boldsymbol{\theta}_i = \left[(\theta_i)_t \right]_{T\times 1}; i,(t)=1,2,\ldots,15,(T)$, $(\theta_i)_{(t:T)}$ represents the tth order statistic, and $\begin{pmatrix} a \\ b \end{pmatrix}$ represents the binomial coefficient.

2.3 Data Preparation

Data preparation, also known as data munging, is crucial for crafting high-quality datasets for modeling. This process included outlier detection and correction, variable transformation, data integration, formatting, splitting, and feature engineering. In this study, mild and extreme outliers are identified utilizing the 1.5 interquartile range (IQR) and the 3 IQR rules, respectively. Mathematically, the lower inner fence (LIF) and upper inner fence (UIF) for the 1.5 IQR rule, and lower outer fence (LOF) and the upper outer fence (UOF) are denoted in equations (5)-(8), respectively.

$$\mathrm{LIF}_{\boldsymbol{\theta}_i} = Q_{1\boldsymbol{\theta}_i} - 1.5\left(Q_{3\boldsymbol{\theta}_i} - Q_{1\boldsymbol{\theta}_i}\right) \tag{5}$$

$$\mathrm{UIF}_{\boldsymbol{\theta}_i} = Q_{3\boldsymbol{\theta}_i} + 1.5\left(Q_{3\boldsymbol{\theta}_i} - Q_{1\boldsymbol{\theta}_i}\right) \tag{6}$$

$$\mathrm{LOF}_{\boldsymbol{\theta}_i} = Q_{1\boldsymbol{\theta}_i} - 3\left(Q_{3\boldsymbol{\theta}_i} - Q_{1\boldsymbol{\theta}_i}\right) \tag{7}$$

$$\mathrm{UOF}_{\boldsymbol{\theta}_i} = Q_{3\boldsymbol{\theta}_i} + 3\left(Q_{3\boldsymbol{\theta}_i} - Q_{1\boldsymbol{\theta}_i}\right) \tag{8}$$

where $Q_{1\boldsymbol{\theta}_i}$ and $Q_{3\boldsymbol{\theta}_i}$ represent the first and third quartiles corresponding to each observation falling below LIF and exceeding UIF, or falling below LOF and exceeding UOF, are identified as mild and extreme outliers, respectively. However, in this study, extreme outlier correction is not performed due to the proposed modified stacked ensemble AI-based predictive algorithm's robustness to the outliers. The primary goal of outlier detection is to evaluate the appropriateness of the numerical EDA measures employed. In contrast, variable transformation is not imposed in this study. Data integration and formatting have been employed to integrate all $\boldsymbol{\theta}_i$ into a cohesive dataset and save it in CSV format. These steps are essential for data splitting, feature engineering, modeling, and evaluation.

Specifically, this study explored the effectiveness of different training-to-test ratios (60:40, 70:30, 80:20, and 90:10) in splitting the integrated dataset to identify the optimal sample size for training the proposed AI-based predictive algorithm. Additionally, this study employed the hybrid stepwise automatic wrapper and Student's *t*-test feature selection (SAWFS) method with the MLR algorithm, utilizing the Akaike Information Criterion (AIC) for evaluation metrics. Since not all selected $\boldsymbol{\theta}_i$; $i \neq 1$ utilizing the SAWFS method are statistically significant, further steps are taken to remove the insignificant $\boldsymbol{\theta}_i$ corresponding to the MLR algorithm in a parsimony feature set. This feature selection method is validated in a previous study related to Malaysia [18], aiming to improve the algorithm's predictive power and interpretability. In summary, these approaches in data preparation offered a robust methodology for algorithm development, ensuring appropriate training data size, relevant feature selection, and reliable evaluation.

2.4 Modeling

The primary data mining objective of this study is to propose a modified stacked ensemble AI-based predictive algorithm. The methodology involved a modeling phase where machine learning algorithms are trained on a split training dataset, and their predictive performance is evaluated utilizing a split test dataset. For this study, the base algorithm selected is the MLR algorithm [18], which plays a crucial role in feature engineering associated with the SAWFS method. Meta-algorithms such as ε-SVR and ν-SVR [25] algorithms are employed, with the principal differences between them lying in their insensitive loss function. The proposed modified stacked ensemble AI-based predictive algorithms include two variants: the modified stacked ensemble MLR-ε-SVR and the modified stacked ensemble MLR-ν-SVR. These variants depart from the conventional approach of training numerous machine learning algorithms and integrating their values utilizing a meta-algorithm. The conventional approach significantly increases algorithm complexity and restricts the interpretability of determinants in practical applications, posing challenges in evaluating the statistical significance of these determinants. These limitations contradicted the business objective of this study.

Specifically, the MLR algorithm is utilized for feature engineering and initial prediction of a 5-year future $\boldsymbol{\theta}_1$, with the resulting statistically significant feature set utilized to train multivariable ε-SVR and ν-SVR algorithms with a linear kernel function. The selection of the linear kernel function is informed by preliminary analysis indicating the linear association between $\boldsymbol{\theta}_1$ against $\boldsymbol{\theta}_i; i = 2,3,\ldots,15$. Predictions from the MLR algorithm are updated utilizing the superior modified stacked ensemble AI-based predictive algorithm identified through the modified Taguchi-based VIseKriterijumska Optimizacija I Kompromisno Resenje (Taguchi-based VIKOR) multi-criteria decision-making algorithm, enhancing the reliability and validity of predictions. The principal advantage of the proposed modified stacked ensemble AI-based predictive algorithm lies in its independence from pre-defined OLS assumptions, low computational cost, compatibility with mid-end spec computers, and robustness to the outliers. Despite these advantages, further investigation of pre-defined OLS assumptions is warranted to enhance prediction reliability and validity.

2.5 Evaluation

This study employed both internal and hold-out cross-validation methods to evaluate prediction performance, utilizing the split test set for a comprehensive evaluation. Diverging from conventional machine learning methods, which frequently relied solely on graphical representation, known for their subjectivity and lack of robustness, this study adopted a more rigorous approach for evaluation. To facilitate optimal decision-making in selecting the superior modified stacked ensemble AI-based predictive algorithm, a modified Taguchi-based VIKOR multi-criteria decision-making algorithm is employed. The evaluation criteria included RMSE, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the average absolute difference of the GoF measures between the training and test sets (AD). These criteria are mathematically expressed as equations (9)-(12).

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\theta_{1t} - \theta_{1t}^{*}\right)^2} \tag{9}$$

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}\left|\theta_{1t} - \theta_{1t}^{*}\right| \tag{10}$$

$$\text{MAPE} = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{\theta_{1t} - \theta_{1t}^{*}}{\theta_{1t}}\right| \tag{11}$$

$$\text{AD} = \left|\frac{1}{3}\left(\text{RMSE}_{\text{Training}} + \text{MAE}_{\text{Training}} + \text{MAPE}_{\text{Training}} - \text{RMSE}_{\text{Test}} - \text{MAE}_{\text{Test}} - \text{MAPE}_{\text{Test}}\right)\right| \tag{12}$$

where θ_{1t}^{*} represents the predicted value of θ_{1t}.

To determine the superior modified stacked ensemble AI-based predictive algorithm, this study employed a modified Taguchi-based VIKOR multi-criteria decision-making algorithm. This decision is prompted by discrepancies in the GoF measures between the training and test sets. Specifically, all GoF measures are transformed into a normalized matrix, $\boldsymbol{\Omega} = \left[\psi_{pq}\right]_{P\times Q}$; $p,(q) = 1, 2,\ldots, P,(Q)$, composed of P modified stacked ensemble AI-based predictive algorithms (alternatives) and Q GoF measures (criteria). Subsequently, ψ_{pq} is converted into Taguchi Design's signal-to-noise ratio (SNR) values, as denoted in equation (13).

$$\text{SNR}_{pq} = -10\log\left(\psi_{pq}\right) \tag{13}$$

The superiority of the modified stacked ensemble AI-based predictive algorithms is determined by ranking them based on the reverse direction of the conventional VIKOR multi-criteria decision-making algorithm. Mathematically, the modified VIKOR multi-criteria decision-making function (η_p) can be expressed as

$$\eta_p = \text{rev}\left(\kappa\left(\frac{S_p - \min\limits_p\left\{S_p\right\}}{\max\limits_p\left\{S_p\right\} - \min\limits_p\left\{S_p\right\}}\right) + \left(1-\kappa\right)\left(\frac{R_p - \min\limits_p\left\{R_p\right\}}{\max\limits_p\left\{R_p\right\} - \min\limits_p\left\{R_p\right\}}\right)\right) \tag{14}$$

where $\text{rev}(\cdot)$ represents the reverse function, $S_p = \sum_{q=1}^{Q} \left\{ \frac{\hat{\sigma}_q^2 \left(\lambda_q^+ - \psi_{pq} \right)}{\lambda_q^+ - \lambda_q^-} \right\}$, and $R_p = \max_q \left\{ \frac{\hat{\sigma}_q^2 \left(\lambda_q^+ - \psi_{pq} \right)}{\lambda_q^+ - \lambda_q^-} \right\}$ represents the utility and regret measures for the predictive algorithms with the GoF measures function, $\lambda_q^+ = \max_p \left\{ \psi_{pq} \right\}$, and $\lambda_q^- = \min_p \left\{ \psi_{pq} \right\}$, and $\hat{\sigma}_q^2 = \frac{\frac{1}{P-1} \sum_{p=1}^{P} \left\{ \psi_{pq} - \bar{\psi}_{pq} \right\}^2}{\sum_{q=1}^{Q} \left\{ \frac{1}{P-1} \sum_{p=1}^{P} \left\{ \psi_{pq} - \bar{\psi}_{pq} \right\}^2 \right\}}; \sum_{q=1}^{Q} \sigma_q^2 = 1, \bar{\psi}_{pq} = \frac{1}{P} \sum_{p=1}^{P} \left\{ \psi_{pq} \right\}$ represents the weight for each GoF measure. This study was set $\kappa = 0.5$ due to a lack of prior knowledge. The decision to employ rank reversal is primarily driven by the resulting irrational ranking among the predictive algorithms observed during a thorough check conducted via the resulting GoF measures. Rank reversal is a widely applied method in numerous multi-criteria decision-making algorithms [35]. Additionally, this study employed the MLR algorithm as a benchmark comparison. However, ε -SVR and ν -SVR algorithms are excluded from benchmark comparisons due to prevalent misconceptions in the literature regarding machine learning algorithms' forecasting capabilities [36]-[37], which are distinct from the predictive performance evaluation.

2.6 Deployment

The superior modified stacked ensemble AI-based predictive algorithm, identified during the evaluation phase is deployed to the forecast 5-year future $\boldsymbol{\theta}_1$, significantly enhancing decision-making in the agricultural realm. Additionally, there's potential for the predictive algorithm to be published as an article, aiming to solicit valuable feedback from reviewers and contribute to scientific knowledge in the field, currently at a Technology Readiness Level of 3 (TRL3). Moreover, the superior modified stacked ensemble AI-based predictive algorithm holds promise for adaption into an interactive dashboard, which could facilitate easy accessibility and insight generation. Furthermore, it is envisioned to be deployed as a technology at TRL4 by integrating with Arduino Integrated Development Environment (IDE) and Internet of Things (IoT) technology. This integration enables real-time monitoring and decision support in agriculture, with applications such as ensuring food security. In summary, this integration demonstrates the predictive algorithm's scalability and its potential to significantly impact the agricultural industry.

3. Analysis Results and Discussion

In this section, all analysis results are meticulously examined utilizing free and open-source software such as R statistical software, supplemented by Microsoft Excel to minimize costs. The analyses are conducted on a mid-end spec computer (Intel(R) Core (TM) i5-10210U CPU @ 1.60GHz, 4 Core(s), 8 Logical Processor(s)). To maintain clarity and coherence, the presentation of analyses, in this section aligned with the phases of the CRISP-DM data science framework. Specifically, the analysis results of the initial phases (Business Understanding, Data Understanding, and Data Preparation) are discussed in Section 3.1, while those associated with the concluding phases (Modeling, Evaluation, and Deployment) are elaborated in Section 3.2.

3.1 Business Understanding, Data Understanding, and Data Preparation

In Table 1, the analysis results of EDA utilizing the first four L-moments and detecting outliers are outlined. As presented earlier, L-moments are selected for their robustness in summarizing statistics, particularly in the presence of outliers and small sample sizes. Table 1 illustrated that some variables employed in this study comprised mild and extreme outliers, highlighting the suitability of L-moments over classical statistical moments for effectively characterizing numerical summary statistics. Additionally, it's worth noting that the analysis results of the nine correlograms associated with the Pearson correlation coefficient corresponding to each Southeast Asia nation have revealed that merely a limited number of determinants exhibited substantial correlation ($|\hat{\rho}| > 0.7$) across each Southeast Asia nation. However, due to space limitations within this article, this study is unable to include these analysis results. Moreover, high correlation determinants are not excluded from this section, as the provided figures do not adequately depict multicollinearity.

Among the nine selected nations, MMR, VNM, and THA ranked among the top ten leading rice exporters globally in 2019 [38]. Consequently, Table 1 demonstrates that the annual average of $\boldsymbol{\theta}_1$ for these nations is significantly higher compared to other Southeast Asia nations, with the exception of IDN. Despite not being the leading exporter of $\boldsymbol{\theta}_1$ globally, IDN prioritizes $\boldsymbol{\theta}_1$ for domestic consumption due to its self-sufficiency ratio

(SSR) and demand ratio of $\boldsymbol{\theta}_1$, which achieved 90% from 2019 to 2021 [39]. Moreover, Table 1 revealed significant fluctuation in the average of certain determinants across Southeast Asia nations, as supported by τ_2 values (bold) exceeding acceptable statistical thresholds [40]. These fluctuations may be attributed to various factors, including geographical considerations, national and international economic strategies and policies, and agriculture policies adopted by policymakers. These variabilities could impact the $\boldsymbol{\theta}_1$ in practice. In developing the proposed modified stacked ensemble AI-based predictive algorithms, it's noted that most dataset fluctuations are insignificant, except for KHM as detailed in Section 3.2.

In statistical theory, τ_3 and τ_4 are frequently utilized to evaluate the shape of variable distributions based on a rule of thumb. However, relying solely on this technique may lead to subjective and suboptimal decisions without proper statistical evidence. Therefore, this study employed the Shapiro-Wilk normality test, as detailed in Table 1. The analysis revealed that not all variables across the Southeast Asia nations considered in this study followed a normal distribution. However, it is important to note that the normality of the dataset is not a prerequisite for developing the proposed modified stacked ensemble AI-based predictive algorithms. These predictive algorithms are not bound by pre-defined statistical assumptions, such as the requirements for residuals to be independently and identically normally distributed. To address feature selection, this study utilized the SAWFS method, as outlined in Section 2. However, the results of the statistical feature selection are not presented in this section. Instead, the analysis discussed in a subsequent section alongside the modeling analysis results for a comprehensive discussion.

Table 1 *Unveiling insights: harnessing L-moments, 1.5 IQR, and 3 IQR rule in EDA analysis*

Country	Variables	Descriptive statistics				Outlier detection	
		τ_1	τ_2	τ_3	τ_4	**1.5 IQR**	**3 IQR**
KHM	$\boldsymbol{\theta}_1$ *	4031166.0000	**0.3946**	0.3144	0.1139	**Yes**	No
	$\boldsymbol{\theta}_2$	31.6702	0.0076	0.0288	0.0882	No	No
	$\boldsymbol{\theta}_3$	1834.1910	0.0472	0.0143	0.1099	No	No
	$\boldsymbol{\theta}_4$ *	0.1871	**0.5602**	0.4519	0.2735	**Yes**	**Yes**
	$\boldsymbol{\theta}_5$ *	197.6814	**0.5117**	0.3248	0.1425	**Yes**	No
	$\boldsymbol{\theta}_6$ *	3118881.0000	**0.4443**	0.4849	0.1681	**Yes**	**Yes**
	$\boldsymbol{\theta}_7$ *	1529.7630	0.0252	-0.1289	0.1995	**Yes**	No
	$\boldsymbol{\theta}_8$ *	55508.4700	**0.5045**	0.2363	0.1143	**Yes**	No
	$\boldsymbol{\theta}_9$	NA	NA	NA	NA	NA	NA
	$\boldsymbol{\theta}_{10}$ *	87.7952	0.1437	0.1198	-0.0600	No	No
	$\boldsymbol{\theta}_{11}$ *	9896397.0000	0.2038	0.1487	-0.0820	No	No
	$\boldsymbol{\theta}_{12}$	16.2479	0.1858	-0.0868	0.1262	**Yes**	No
	$\boldsymbol{\theta}_{13}$ *	78.5225	0.1398	-0.2104	-0.0709	No	No
	$\boldsymbol{\theta}_{14}$	1996421.0000	0.1905	-0.0194	0.1107	No	No
	$\boldsymbol{\theta}_{15}$ *	0.0698	**0.5209**	0.4517	0.1389	**Yes**	No
LAO	$\boldsymbol{\theta}_1$ *	1767986.0000	**0.3388**	0.2551	0.0366	No	No
	$\boldsymbol{\theta}_2$	28.7986	0.0087	0.0739	0.1497	**Yes**	No
	$\boldsymbol{\theta}_3$	1830.1770	0.0559	-0.0256	0.0688	No	No
	$\boldsymbol{\theta}_4$ *	0.3631	**0.6681**	0.6875	0.4830	**Yes**	**Yes**
	$\boldsymbol{\theta}_5$ *	197.6814	**0.5117**	0.3248	0.1425	**Yes**	No
	$\boldsymbol{\theta}_6$ *	1428288.0000	**0.3949**	0.4321	0.1485	**Yes**	No
	$\boldsymbol{\theta}_7$ *	1470.4880	0.0332	0.2808	0.0210	No	No
	$\boldsymbol{\theta}_8$ *	50322.0300	**0.4817**	0.2726	0.0694	No	No
	$\boldsymbol{\theta}_9$	NA	NA	NA	NA	NA	NA
	$\boldsymbol{\theta}_{10}$ *	71.2219	0.1666	0.0092	-0.0616	No	No
	$\boldsymbol{\theta}_{11}$ *	4449597.0000	0.2047	0.0576	-0.0378	No	No
	$\boldsymbol{\theta}_{12}$ *	18.4587	0.2694	0.1875	-0.0106	No	No
	$\boldsymbol{\theta}_{13}$ *	55.3222	0.2081	0.2835	0.0683	No	No
	$\boldsymbol{\theta}_{14}$ *	717618.2000	0.1038	0.1341	0.0610	No	No
	$\boldsymbol{\theta}_{15}$ *	0.0347	**0.3515**	0.1643	0.2271	**Yes**	**Yes**
MMR	$\boldsymbol{\theta}_1$ *	16928990.0000	0.2694	0.1458	-0.0131	No	No
	$\boldsymbol{\theta}_2$	28.9663	0.0062	0.1057	0.1320	No	No
	$\boldsymbol{\theta}_3$	2036.2990	0.0454	-0.0425	0.1118	No	No

	θ_4 *	0.2022	0.2587	0.4281	0.3057	**Yes**	**Yes**
	θ_5 *	197.6814	**0.5117**	0.3248	0.1425	**Yes**	No
	θ_6 *	12105830.0000	**0.3487**	0.2399	0.0574	No	No
	θ_7 *	1141.1500	0.0566	-0.0640	-0.0801	No	No
	θ_8 *	4949.1530	**0.8375**	0.6992	0.4161	**Yes**	**Yes**
	θ_9 *	13.3689	**0.5753**	0.2045	0.0743	No	No
	θ_{10} *	85.9892	0.1100	-0.2273	-0.0552	No	No
	θ_{11} *	38938500.0000	0.1420	-0.0686	-0.0241	No	No
	θ_{12}	25.5398	0.0655	0.0048	0.0774	No	No
	θ_{13} *	83.4187	0.0551	0.3756	0.0332	No	No
	θ_{14} *	5666431.0000	0.1144	0.2593	-0.0050	No	No
	θ_{15} *	0.1076	**0.5539**	0.4627	0.2175	**Yes**	**Yes**
PHL	θ_1 *	10380770.0000	0.2701	0.1227	0.0074	No	No
	θ_2	30.3339	0.0059	-0.0039	0.1442	No	No
	θ_3	2551.1740	0.0718	0.0713	0.1578	**Yes**	No
	θ_4	0.7678	0.1555	0.0117	0.2112	**Yes**	No
	θ_5 *	197.6814	**0.5117**	0.3248	0.1425	**Yes**	No
	θ_6 *	8582085.0000	**0.3555**	0.3062	0.0853	**Yes**	No
	θ_7 *	956.6432	0.0924	0.1252	0.0341	No	No
	θ_8 *	769372.9000	**0.6341**	0.4095	0.1541	**Yes**	**Yes**
	θ_9 *	8.6525	**0.4449**	0.3982	0.2770	**Yes**	**Yes**
	θ_{10} *	83.4266	0.1293	-0.2067	0.0376	No	No
	θ_{11} *	64792500.0000	0.2195	0.0830	-0.0119	No	No
	θ_{12} *	41.3441	0.0779	-0.2676	-0.0689	No	No
	θ_{13} *	84.6634	0.0784	-0.1807	0.0990	No	No
	θ_{14} *	3746360.0000	0.0824	0.1971	-0.0056	No	No
	θ_{15} *	1.3322	**0.5208**	0.2646	0.0385	No	No
TLS	θ_1 *	44440.4900	**0.3321**	0.2048	0.1477	**Yes**	No
	θ_2	29.3314	0.0055	-0.0482	0.0833	No	No
	θ_3 *	1305.3780	0.0931	0.1461	0.2423	**Yes**	**Yes**
	θ_4	NA	NA	NA	NA	NA	NA
	θ_5 *	197.6814	**0.5117**	0.3248	0.1425	**Yes**	No
	θ_6 *	66508.4700	0.2304	0.4627	0.2284	**Yes**	**Yes**
	θ_7 *	627.1337	0.0715	-0.1066	0.0663	No	No
	θ_8 *	33169.4900	0.2676	0.3869	0.5528	**Yes**	**Yes**
	θ_9	NA	NA	NA	NA	NA	NA
	θ_{10} *	92.9518	0.0679	-0.1058	0.0115	No	No
	θ_{11} *	797351.4000	0.1688	0.1512	0.0300	No	No
	θ_{12} *	20.4417	0.1801	-0.0041	-0.0188	No	No
	θ_{13} *	69.1668	0.2110	-0.0269	-0.1109	No	No
	θ_{14} *	20042.9800	**0.3023**	0.2298	0.1103	**Yes**	No
	θ_{15} *	0.0384	**0.4135**	0.2774	0.1635	**Yes**	No
VNM	θ_1 *	23332070.0000	**0.3160**	0.1526	-0.0857	No	No
	θ_2	28.4010	0.0072	0.0397	0.1474	**Yes**	No
	θ_3 *	1767.0760	0.0336	0.0068	-0.0149	No	No
	θ_4 *	0.8234	**0.4380**	0.4694	0.1938	**Yes**	**Yes**
	θ_5 *	197.6814	**0.5117**	0.3248	0.1425	**Yes**	No
	θ_6 *	14412930.0000	0.2935	0.2411	0.0179	No	No
	θ_7 *	1430.5600	0.0259	-0.1704	0.2880	**Yes**	**Yes**
	θ_8 *	243271.2000	**0.7188**	0.5111	0.1837	**Yes**	**Yes**
	θ_9	NA	NA	NA	NA	NA	NA
	θ_{10} *	86.8255	0.1149	-0.1042	-0.0537	No	No
	θ_{11} *	65544520.0000	0.1725	-0.0302	-0.0384	No	No
	θ_{12} *	22.9775	0.1426	0.2498	0.0767	No	No
	θ_{13} *	65.6091	0.1549	0.2257	-0.0194	No	No

	θ_{14} *	6275102.0000	0.1045	-0.0118	-0.1009	No	No
	θ_{15} *	0.3518	**0.6373**	0.4484	0.0780	**Yes**	No
IDN	θ_{1} *	39429630.0000	0.2378	-0.1316	-0.0479	No	No
	θ_{2}	30.3678	0.0040	-0.0461	0.0917	No	No
	θ_{3}	2754.4340	0.0454	-0.1054	0.0871	No	No
	θ_{4} *	1.0275	**0.3575**	0.1444	0.0074	No	No
	θ_{5} *	197.6814	**0.5117**	0.3248	0.1425	**Yes**	No
	θ_{6} *	30350390.0000	0.2986	0.1136	0.0574	No	No
	θ_{7} *	1190.5040	0.0656	-0.3104	0.2202	**Yes**	No
	θ_{8} *	1135458.0000	**0.4886**	0.3206	0.1617	**Yes**	No
	θ_{9} *	44.6816	**0.7932**	0.8371	0.7163	**Yes**	**Yes**
	θ_{10} *	88.2896	0.1058	-0.1248	-0.0926	No	No
	θ_{11} *	180676600.0000	0.1761	-0.0098	-0.0137	No	No
	θ_{12} *	32.7439	0.2425	0.0840	-0.0682	No	No
	θ_{13} *	70.1968	0.1645	0.1608	-0.0100	No	No
	θ_{14} *	9984559.0000	0.0936	-0.1569	-0.0491	No	No
	θ_{15} *	0.0320	**0.4874**	0.2111	-0.0247	No	No
MYS	θ_{1}	1970238.0000	0.1240	-0.0674	0.1275	No	No
	θ_{2}	30.2458	0.0066	0.0316	0.0970	No	No
	θ_{3}	2922.7170	0.0553	0.0601	0.0613	No	No
	θ_{4} *	4.0573	**0.3645**	0.0910	-0.0696	No	No
	θ_{5} *	197.6814	**0.5117**	0.3248	0.1425	**Yes**	No
	θ_{6} *	2078695.0000	0.2349	0.3787	0.1581	**Yes**	**Yes**
	θ_{7} *	953.6877	0.1269	0.1522	-0.0473	No	No
	θ_{8} *	643915.3000	**0.3753**	0.3819	0.1611	**Yes**	**Yes**
	θ_{9} *	3.0321	**0.4593**	0.2519	0.2592	**Yes**	**Yes**
	θ_{10}	90.1016	0.0327	-0.0571	0.0341	No	No
	θ_{11} *	18813600.0000	0.2394	0.1026	-0.0308	No	No
	θ_{12} *	51.6514	0.1759	0.0228	-0.0337	No	No
	θ_{13} *	68.9910	0.1639	-0.0729	0.0012	No	No
	θ_{14} *	670000.6000	0.0409	-0.1858	0.2841	**Yes**	**Yes**
	θ_{15} *	0.1884	**0.4416**	0.1971	-0.0123	No	No
THA	θ_{1} *	21863300.0000	0.2114	0.1165	0.0067	No	No
	θ_{2}	31.8581	0.0077	0.0631	0.1292	**Yes**	No
	θ_{3}	1588.9650	0.0523	0.0123	0.1497	**Yes**	No
	θ_{4} *	1.9537	**0.4146**	0.1041	-0.0948	No	No
	θ_{5} *	197.6814	**0.5117**	0.3248	0.1425	**Yes**	No
	θ_{6} *	10517640.0000	0.2158	0.3803	0.1997	**Yes**	**Yes**
	θ_{7} *	1300.5080	0.0943	0.1961	-0.0298	No	No
	θ_{8} *	14796.6100	**0.8186**	0.6498	0.3015	**Yes**	**Yes**
	θ_{9} *	4.2572	**0.5203**	0.3142	0.2727	**Yes**	**Yes**
	θ_{10} *	116.5800	0.1095	0.1116	0.0443	No	No
	θ_{11} *	53013210.0000	0.1507	-0.1094	-0.0265	No	No
	θ_{12} *	31.1928	0.1653	0.1881	0.0732	No	No
	θ_{13} *	82.2344	0.1032	-0.2619	0.0736	No	No
	θ_{14} *	9120229.0000	0.0961	-0.0910	0.1145	No	No
	θ_{15} *	0.8327	**0.4777**	0.1019	-0.1534	No	No

***Note:** "*" represents the corresponding variable a non-normal distribution that was verified utilizing the Shapiro-Wilk normality test, while NA represents the variable that is not available.

3.2 Modeling, Evaluation, and Deployment

This study aimed to develop a modified stacked ensemble MLR-SVR-based algorithm and compare its effectiveness to benchmarks. Table 2 depicts a ranking analysis of proposed AI-based predictive algorithms against benchmarks, particularly focusing on the proposed modified stacked ensemble MLR-SVR-based algorithms across various training-to-test ratios. Results consistently revealed the superiority of the proposed modified stacked ensemble AI-based predictive algorithm, except for PHL, VNM, and THA, where other predictive algorithms performed better. However, upon closer examination, this study found that in PHL, VNM, and THA, the proposed modified stacked ensemble AI-based predictive algorithms outperformed the traditional MLR algorithm due to their flexibility and independence from pre-defined OLS assumptions.

To ensure the reliability and validity of the proposed modified stacked ensemble AI-based predictive algorithms, various tests were conducted, including the Shapiro-Wilks test for normality, the run test for independence, and the Breusch-Pagan test for homoscedasticity. Results verified that, except for TLS, the superior modified stacked ensemble MLR-SVR-based algorithms for each Southeast Asia nation did not violate the OLS assumptions. Importantly, the violation of these assumptions did not compromise the predictive capability of the proposed modified stacked ensemble AI-based predictive algorithms. Moreover, although the GoF measures for the training and test sets could not be included in this article due to limitations, these measures further supported the effectiveness of the proposed modified stacked ensemble AI-based predictive algorithms. Additionally, Table 2 also revealed that the superiority of the proposed modified stacked ensemble AI-based predictive algorithm associated with the optimal training-to-test ratios varied across nations. In simple terms, there was no universal optimal training-to-test ratio for training the proposed modified stacked ensemble AI-based predictive algorithm, as it could be affected by various factors such as the size and complexity of the dataset, the nature of the addressed problems, and the employed predictive algorithms. The finding of this study was consistently valid from the machine learning perspective.

In practice, the determinants included in the superior modified stacked ensemble MLR-SVR-based algorithms across Southeast Asia nations are attributed to geographical determinants, national and international economic strategies and policies, and agriculture policies adopted by policymakers. This practical statement is further authenticated based on the findings presented in Table 3, which elucidate the association between these determinants and their impacts on agricultural outcomes. Particularly, the determinants across Southeast Asia nations that are statistically significantly affected $\boldsymbol{\theta}_1$ are varied. For convincing interpretation, this study further partitioned all the determinants considered into three clustered principles: atmospheric, socioeconomic, and farming practices. This partitioning is necessary to provide a structured framework for analyzing the diverse determinants influencing and to facilitate a deeper understanding of their contribution.

In brief, the findings of this study revealed that atmospheric clustered determinants statistically affected $\boldsymbol{\theta}_1$ merely for MMR, IDN, and THA and vice versa in other low-middle-income and upper-middle-income Southeast Asia nations. Conversely, the analysis results indicated that both socioeconomic and farming practices statistically significantly affected $\boldsymbol{\theta}_1$ for all Southeast Asia nations considered in this study. This comprehensive analysis underscored the multifaceted nature of determinants influencing $\boldsymbol{\theta}_1$ and highlights the importance of considering diverse determinants in agricultural research and policymaking. Consequently, these findings are also consistently valid from the agriculture economy perspective, primarily due to the occurrence of global climate change, the exponential growth of population density, global food inflation, the arrival of Industrial Revolution 4.0 (IR4.0), and the revised national and international economy policies principally focusing on the green economy. These overreaching trends underscored the need for adaptive and sustainable approaches to agricultural development and emphasized the importance of integrating environmental, economic, and technological considerations into agricultural policies and practices.

Building upon these insights, this study deployed the superior modified stacked ensemble MLR-SVR-based algorithm to forecast the 5-year future trends of $\boldsymbol{\theta}_1$ for each Southeast Asia nation, as illustrated in Figure 2. The analysis focused on presenting forecasted results for both the top-ranked (Figure 2(a)) and bottom-ranked (Figures 2(b)-2(c)) nations, determined by the modified Taguchi-based VIKOR multi-criteria decision-making algorithm. In Figure 2(a), the proposed modified stacked ensemble MLR-SVR-based algorithm for LAO emerged as the most suitable among the nine superior modified stacked ensemble AI-based predictive algorithms for Southeast Asia nations. The forecasted trends of $\boldsymbol{\theta}_1$ showed a consistent increase from 2020 to 2024, aligning with forecasts by the United States Department of Agriculture [41] and supporting the implementation of Lao policymakers' Agricultural Development Strategy, which prioritized $\boldsymbol{\theta}_1$ and export by 2025 [41].

Similarly, Figure 2(c) indicated increasing trends in the 5-year futures of $\boldsymbol{\theta}_1$ for KHM. This finding is consistent with the continuous growth in $\boldsymbol{\theta}_1$ in KHM observed from 2012 to 2020 [42], and the targeted increase set by the Cambodia Rice Federation in collaboration with the Ministry of Commerce up to 2025 to assess new market [43]. However, a decline in the forecast between 2019 and 2020 suggested the limitations of the proposed modified stacked ensemble AI-based predictive algorithm. Therefore, a decline in the forecast suggested limitations in the proposed modified stacked ensemble AI-based predictive algorithm, prompting further exploration of more suitable kernels such as the polynomial kernel. To further validate these findings, this study also presented forecasted results based on the second-bottom-ranked algorithm for MYS in Figure 2(b). The forecast indicated an increasing trend in $\boldsymbol{\theta}_1$ for the next 5 years, supported by mild evidence from statistics provided by the Department of Statistics Malaysia [44]. However, a decrease that occurred in 2022 was noted in 2023 [45], primarily attributed to uncontrollable determinants such as the insufficient supply of basic paddy seeds and prevailing farming practices [46]. These aspects are influenced by broader agricultural policies and conditions beyond the determinants considered in this study. While there was uncertainty regarding formal statistics on paddy and $\boldsymbol{\theta}_1$ beyond 2022, the trend suggested a potential continuation of growth. This forecasted result is consistent with Malaysia's target to achieve an SSR of 73.8% in 2022 and 80% by 2030 [46], demonstrating alignment with national agriculture goals.

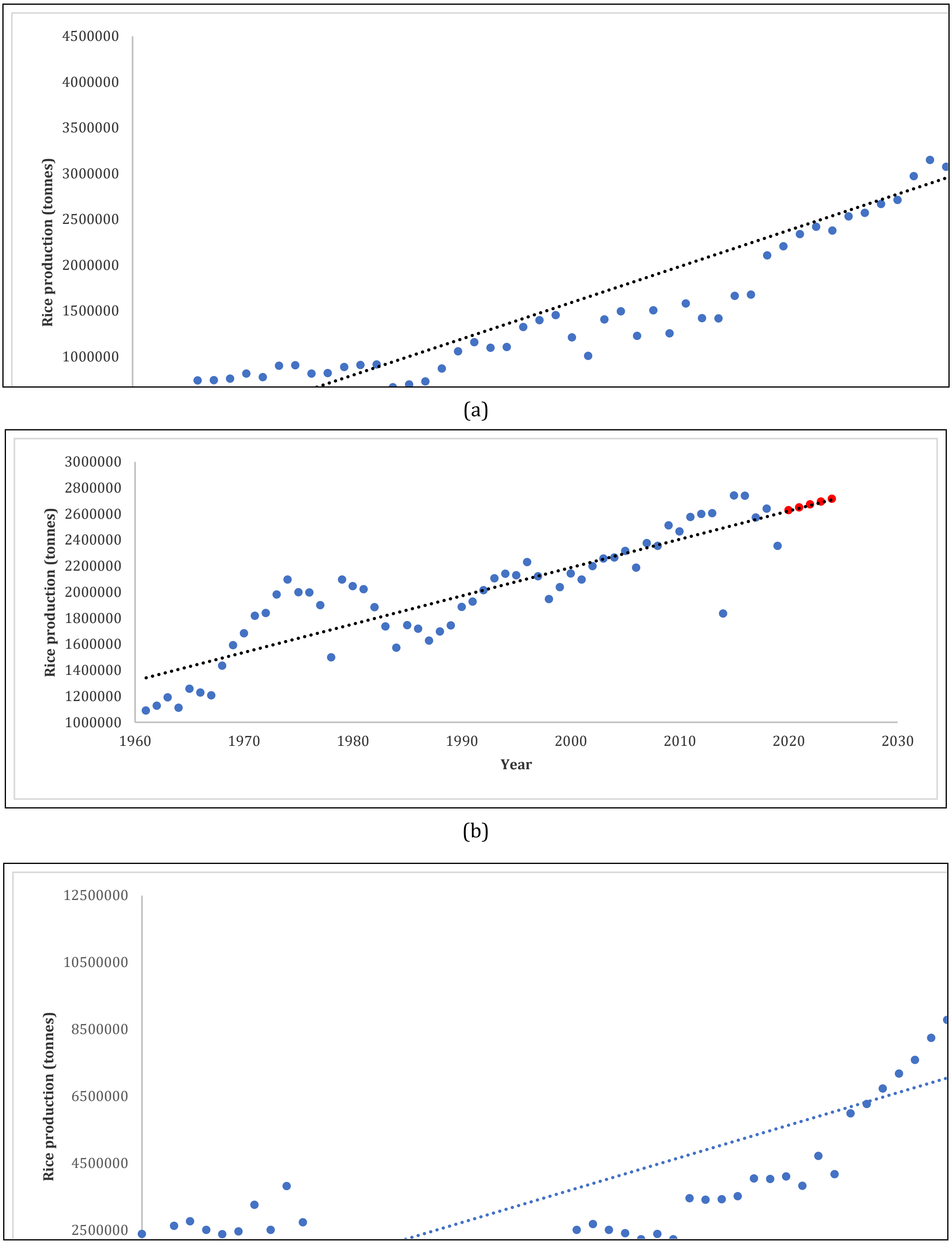
This article's findings were crucial for policymakers, as they provided valuable insights into the impact of atmospheric clustered determinants on $\boldsymbol{\theta}_1$ and served as an early warning regarding climate change's effect on this critical staple food. Moreover, these findings highlighted the significant roles played by socioeconomic and farming practices clustered determinants in impacting both national and regional $\boldsymbol{\theta}_1$, offering valuable guidance to policymakers and farmers, especially in the context of technological advancements and Industrial Revolution 4.0 (IR4.0). Aligned with several regional SDGs, including SDG1, SDG2, SDG3, and SDG8, this study emphasized rice's pivotal role as a staple food in Southeast Asia. Agriculture emerged as a key sector capable of generating job opportunities, increasing household income, promoting women's participation in the labor market, and enhancing food security. In brief, fostering economic growth through agriculture, while considering atmospheric, socioeconomic, and farming practices clustered determinants, is essential for long-term prosperity and sustainable development. This approach propelled low-middle-income and upper-middle-income Southeast Asia nations toward achieving high-income countries.

Table 2 *Empowering predictive algorithms: exploring superior predictive algorithm through modified Taguchi-based VIKOR multi-criteria decision-making analysis*

Algorithm	Ratio	Rank								
		Low-middle-income						Upper-middle-income		
		KHM	LAO	MMR	PHL	TLS	VNM	IDN	MYS	THA
MLR	60:40	12	11	2	9	2	6	12	12	11
MLR	70:30	9	3	4	1	4	1	9	10	4
MLR	80:20	6	5	5	5	9	4	3	5	1
MLR	90:10	4	9	6	11	10	3	5	4	9
MLR- ε -SVR	60:40	10	12	1	7	3	5	10	11	12
MLR- ε -SVR	70:30	7	4	7	3	6	8	7	9	6
MLR- ε -SVR	80:20	5	6	11	6	8	12	2	8	2
MLR- ε -SVR	90:10	3	8	10	10	12	2	8	6	7
MLR- ν -SVR	60:40	11	10	3	8	1	7	11	1	8
MLR- ν -SVR	70:30	8	1	8	2	5	10	4	2	5
MLR- ν -SVR	80:20	2	2	12	4	7	11	1	7	3
MLR- ν -SVR	90:10	1	7	9	12	11	9	6	3	10

Table 3 *Unveiling the key factors: statistically significant determinants in superior AI-based predictive algorithms across Southeast Asia nations*

Nation	Algorithm	Ratio	Determinant													
			Atmospheric			Socioeconomic								Farming practices		
			$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$	$\boldsymbol{\theta}_5$	$\boldsymbol{\theta}_6$	$\boldsymbol{\theta}_7$	$\boldsymbol{\theta}_8$	$\boldsymbol{\theta}_9$	$\boldsymbol{\theta}_{10}$	$\boldsymbol{\theta}_{11}$	$\boldsymbol{\theta}_{12}$	$\boldsymbol{\theta}_{13}$	$\boldsymbol{\theta}_{14}$	$\boldsymbol{\theta}_{15}$
KHM	MLR- ν -SVR	90:10					√			NA		√		√	√	√
LAO	MLR- ν -SVR	70:30					√	√		NA			√		√	√
MMR	MLR- ε -SVR	60:40	√			√	√	√			√		√		√	
PHL	MLR- ν -SVR	70:30					√		√					√	√	
TLS	MLR- ν -SVR	70:30			NA		√		√	NA					√	√
VNM	MLR- ε -SVR	90:10				√		√				√	√		√	√
IDN	MLR- ν -SVR	80:20			√		√		√		√			√	√	√
MYS	MLR- ν -SVR	60:40				√		√					√		√	
THA	MLR- ε -SVR	80:20		√		√	√	√					√	√	√	



(a)

(b)

(c)

Fig. 2 *Unveiling future trends: deployment of top-ranked and AI-based predictive algorithms in 5-year* $\boldsymbol{\theta}_1$ *forecasts for (a) LAO, (b) MYS, and (c) KHM*

4. Conclusions

The principal objective of this study is to propose an innovative modified stacked ensemble MLR-SVR-based algorithms for $\boldsymbol{\theta}_1$ in six low-middle-income (KHM, LAO, MMR, PHL, TLS, and VNM) and three upper-middle-income (IDN, MYS, and THA) nations in Southeast Asia, utilizing the CRISP-DM data science framework. The MLR algorithm served as a benchmark for comparison. A key advantage of the proposed modified stacked ensemble AI-based predictive algorithm is its ability to bypass the pre-defined assumptions of MLR algorithms while retaining interpretability for statistically significant determinants. To evaluate the efficiency of the proposed modified stacked ensemble AI-based predictive algorithms, annual time-series datasets from 1961 to 2019, consisting of one endogenous variable and 14 potential determinants, were employed.

Due to discrepancies in GoF measures between the training and test sets, the modified Taguchi-based VIKOR multi-criteria decision-making algorithm was applied to evaluate the superiority of the AI-based predictive algorithms. The results indicated that no single AI-based predictive algorithm universally fit all nations' datasets across varying training-to-test ratios (60:40, 70:30, 80:20, and 90:10). However, the proposed modified stacked ensemble MLR-SVR-based algorithm performed well for most low-middle-income and upper-middle-income nations, except for PHL, VNM, and THA, where the MLR algorithm was found superior. Despite this, this study concluded that the proposed modified stacked ensemble MLR-SVR-based algorithm was superior for PHL, VNM, and THA based on its lower AD measures.

Additionally, socioeconomic and farming practices were identified as key determinants affecting $\boldsymbol{\theta}_1$ in Southeast Asia, except for MMR, IDN, and THA. These nations, $\boldsymbol{\theta}_1$ is also affected by atmospheric clustered determinants, likely due to the diverse climate types across the region. Specifically, the analysis showed that atmospheric conditions statistically affected $\boldsymbol{\theta}_1$ in these nations, alongside socioeconomic and farming practices clustered determinants such as international trade, human resource economics, and agriculture resources and technologies. However, the variability in determinants (Table 3) across Southeast Asia nations can be attributed to geographical determinants, national and international economic strategies and policies, and agriculture policies adopted by policymakers. Future research could focus on developing predictive algorithms based on the geographical determinants to gain more precise insights into determinants of $\boldsymbol{\theta}_1$ in the region.

Acknowledgments

The authors extend their sincere appreciation to the founders and associations of Climate Change Knowledge Portal (CCKP) and Our World in Data (OWID) for providing the dataset utilized in this study. Additionally, gratitude is extended to Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) for providing financial support such as the UMP Fundamental Research Scheme [Ref. No.: RDU220393]. Special thanks are also given to the reviewers for their constructive comments, which contributed to the improvement of the study's quality.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors utilized ChatGPT to improve the readability and language of this work. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Author Contribution

The authors confirm their contribution to the paper as follows: ***Conceptualization:*** *Zun Liang Chuan; Tham Ren Sheng;* ***Methodology:*** *Zun Liang Chuan;* ***Software:*** *Zun Liang Chuan; Tham Ren Sheng; Tan Check Cheng; Abraham Lim Bing Sern; David Lau King Luen;* ***Validation:*** *Zun Liang Chuan;* ***Formal Analysis:*** *Zun Liang Chuan; Tham Ren Sheng; Tan Check Cheng; Abraham Lim Bing Sern; David Lau King Luen;* ***Writing-Original Draft:*** *Zun Liang Chuan;* ***Writing-Review & Editing:*** *Tham Ren Sheng; Tan Check Cheng; Abraham Lim Bing Sern; David Lau King Luen;* ***Visualization****: Tham Ren Sheng; Tan Check Cheng; Abraham Lim Bing Sern; David Lau King Luen;* ***Project Administration:*** *Zun Liang Chuan; Tham Ren Sheng;* ***Funding Acquisition:*** *Zun Liang Chuan. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] M. S. Gunaratne, R. B. R. Firdaus, and S. I. Rathnasooriya, "Climate change and food security in Sri Lanka: towards food sovereignty," Humanities & Social Sciences Communications, vol. 8, p. 229, 2021, doi: 10.1057/s41599-021-00917-4.

[2] N. K. Fukagawa and L. H. Ziska, "Rice: importance for global nutrition," Journal of Nutritional Science and Vitaminology, vol. 65, Supplement, pp. S2-S3, 2019, doi: 10.3177/jnsv.65.S2.

[3] Food and Agriculture Organization of the United Nations, "FAOSTAT (Data): Suite of food security indicators," Aug. 23, 2023. [Online]. Available: https://www.fao.org/faostat/en/#data/FS

[4] 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, 2021, doi: 10.3390/agriculture11060569.

[5] Z. L. Chuan, S.-F. Fam, Q. H. Lee, J. S. Kok, and M. N. B. M. Azam, "Modelling the impacts of climate change and air pollutants on the agricultural production yields in Malaysia using random-effects error components regression model," Data Analytics and Applied Mathematics, vol. 3, no. 2, pp. 1-12, 2022, doi: 10.15282/daam.v3i2.7755.

[6] J. Duasa and N. A. Mohd-Radzman, "The impact of climate change on ASEAN rice production in short and long-run," Journal of Tourism, Hospitality and Environment Management, vol. 7, no. 29, pp. 94-111, 2022, doi: 10.35631/JTHEM.729008.

[7] T. T. Nguyen, M. H. Do, and D. Rahut, "Shock, risk attitude and rice farming: evidence from panel data for Thailand," Environmental Challenges, vol. 6, p. 100430, 2022, doi: 10.1016/j.envc.2021.100430.

[8] S. Yohandoko and Supriyanto, "Panel data analysis on rice (paddy) production in Indonesia 2018-2021," International Journal of Mathematics, Statistics, and Computing, vol. 1, no. 3, pp. 44-53, 2023.

[9] J. M. G. Dait, "A panel data study on factors affecting rice production in the Philippines," Universal Journal of Agricultural Research, vol. 11, no. 3, pp. 547-557, 2023, doi: 10.13189/ujar.2023.110305.

[10] N. Koide, A. W. Robertson, A. V. M. Ines, J.-H. Qian, D. G. DeWitt, and A. Lucero, "Prediction of rice production in the Philippines using seasonal climate forecasts," Journal of Applied Meteorology and Climatology, vol. 52, no. 3, pp. 552-569, 2013, doi: 10.1175/JAMC-D-11-0254.1.

[11] A. Bashir and S. Yuliana, "Identifying factors influencing rice production and consumption in Indonesia," Jurnal Ekonomi Pembangunan: Kajian Masalah Ekonomi dan Pembangunan, vol. 19, no. 2, pp. 172-185, 2018, doi: 10.23917/jep.v19i2.5939.

[12] N. K. Win, K. K. Win, C. C. San, and N. N. Htwe, "Factors affecting farmers' attitudes on hybrid rice production in Nay Pyi Taw area, Myanmar," International Journal of Environmental and Rural Development, vol. 9, no. 2, pp. 88-94, 2018.

[13] N. Idalisa, M. Rivaie, N. H. Fadhilah, N. Atikah, A. Shahida, and N. H. M. Noh, "Multiple linear regression model of rice production using conjugate gradient methods," Matematika: Malaysian Journal of Industrial and Applied Mathematics, vol. 35, no. 2, pp. 229-235, 2019, doi: 10.11113/matematika.v35.n2.1180.

[14] N. M. Roslan, W. L. Shinyie, and S. S. Ling, "Modelling high dimensional paddy production data using Copulas," Pertanika Journal of Science & Technology, vol. 29, no. 1, pp. 263-284, 2021, doi: 10.47836/pjst.29.1.15.

[15] S. Aprizkiyandari and T. Palupi, "Rice production forecasting model in West Kalimantan with factors are rainfall and harvest area," IOP Conference Series: Earth and Environmental Science, vol. 1177, p. 012015, 2023, doi: 10.1088/1755-1315/1177/1/012015.

[16] K. Saithanu, J. Mekparyup, and P. Phodjanawichaikul, "Forecasting rice yield in the northern Thailand with multiple linear regression model," Global Journal of Pure and Applied Mathematics, vol. 11, no. 4, pp. 2181-2186, 2015.

[17] M. C. V. David, "Rice yield modeling using machine learning algorithms based on environmental and agronomic data of Pampanga River Basin, Philippines," Universal Journal of Agricultural Research, vol. 11, no. 5, pp. 836-848, 2023, doi: 10.13189/ujar.2023.110509.

[18] Z. L. Chuan, D. C. T. Wei, A. S. B. A. Aminuddin, S-F. Fam, and T. L. Ken, "Comparison of multiple linear regression and multiple nonlinear regression models for predicting rice production," AIP Conference Proceedings, vol. 3150, p. 050008, 2024, doi: 10.1063/5.0227872.

[19] X. Meng, M. Liu, and Q. Wu, "Prediction of rice yield via stacked LSTM," International Journal of Agricultural and Environmental Information Systems, vol. 11, no. 1, pp. 86-95, 2020, doi: 10.4018/IJAEIS.2020010105.

[20] H. Mo, Y. Zhang, Y. Liu, and Y. Zheng, "Prediction of rice yield based on LSTM long short term memory network," Journal of Physics: Conference Series, vol. 1952, p. 042033, 2021, doi: 10.1088/1742-6596/1952/4/042033.

[21] M. Geetha, R. C. Suganthe, S. K. Nivetha, R. Anju, R. Anuradha, and J. Haripriya, "A time-series based yield forecasting model using stacked LSTM to predict the yield of paddy in Cauvery Delta zone in Tamilnadu," in Proceedings of the First International Conference on Electrical, Electronics, Information and Communication Technologies, 2022, pp. 1-6, doi: 10.1109/ICEEICT53079.2022.9768441.

[22] P. Sathya and P. Gnanasekaran, "Paddy yield prediction in Tamilnadu Delta region using MLR-LSTM model," Applied Artificial Intelligence, vol. 37, no. 1, p. e2175113, 2023, doi: 10.1080/08839514.2023.2175113.

[23] X. Chang, "Rice yield prediction based on deep learning," in Artificial Intelligence Technologies and Applications, C. Chen, Ed., pp. 490-496, IOS Press, 2024, doi: 10.3233/FAIA231333.

[24] J. Garcia-Arteaga, J. Zambrano-Zambrano, J. Parraga-Alava, and J. Rodas-Silva, "An effective approach for identifying keywords as high-quality filters to get emergency-implicated Twitter Spanish data," Computer Speech & Language, vol. 84, p. 101579, 2024, doi: 10.1016/j.csl.2023.101579.

[25] Z. L. Chuan, O. S. Jie, T. Y. Hin, S. N. S. B. M. Zain, Y. A. B. A. Rashid, and A. N. B. Kamarudin, "Enhancing electricity consumption forecasting in sparse dataset: A simple stacked ensemble approach incorporating simple linear and support vector regression for Malaysia," International Journal of Built Environment & Sustainability, in press.

[26] Z. L. Chuan, N. Japashov, S. K. Yuan, T. W. Qing, and N. Ismail, "Analyzing enrolment patterns: Modified stacked ensemble statistical learning-based approach to educational decision-making," Akademika, vol. 94, no. 2, 2024, doi: 10.17576/akad-2024-9402-13.

[27] C. Z. Liang, A. L. B. Sern, T. C. Cheng, D. L. K. Luen, N. Japashov, and T. E. Hiae, "Empowering Industry 5.0: nurturing STEM tertiary education and careers through Additional Mathematics," in Utilizing Renewable Energy, Technology, and Education for Industry 5.0, S. N. S. Al-Humairi, Ed. IGI Global, 2024, doi: 10.4018/979-8-3693-2814-9.

[28] A. Cheng, "Evaluating Fintech industry's risks: a preliminary analysis based on CRISP-DM framework," Finance Research Letters, vol. 55, Part B, p. 103966, 2023, doi: 10.1016/j.frl.2023.103966.

[29] O. Lohaj, J. Paralič, P. Bednár, Z. Paraličová, and M. Huba, "Unraveling COVID-19 dynamics via machine learning and XAI: investigating variant influence and prognostic classification," Machine Learning & Knowledge Extraction, vol. 5, no. 4, pp. 1266-1281, 2023, doi: 10.3390/make5040064.

[30] C. L. S. Cedric, W. Y. H. Adoni, R. Aworka, J. T. Zoueu, F. K. Mutombo, M. Krichen, and C. L. M. Kimpolo, "Crops yield prediction based on machine learning models: case of West African countries," Smart Agricultural Technology, vol. 2, p. 100049, 2022, doi: 10.1016/j.atech.2022.100049.

[31] A. Alodah, "Towards sustainable water resources management considering climate change in the case of Saudi Arabia," *Sustainability*, vol. 15, no. 20, p. 14674, 2023, doi: 10.3390/su152014674.

[32] Alamo, D. G. Reina, M. Mammarella, and A. Abella, "COVID-19: open-data resources for monitoring, modeling, and forecasting the epidemic," Electronics*,* vol. 9, no. 5, p. 827, 2020, doi: 10.3390/electronics9050827.

[33] R. C. A. Achterbergh, I. McGovern, and M. Haag, "Co-administration of influenza and COVID-19 vaccines: policy review and vaccination coverage trends in the European Union, UK, US, and Canada between 2019 and 2023," Vaccines*,* vol. 12, no. 2, p. 216, 2024, doi: 10.3390/vaccines12020216.

[34] J. R. M. Hosking, "L-moments: analysis and estimation of distributions using linear combinations of order statistics," Journal of the Royal Statistical Society. Series B (Methodological), vol. 52, no. 1, pp. 105-124, 1990.

[35] B. Ceballos, D. A. Pelta, and M. T. Lamata, "Rank reversal and the VIKOR method: an empirical evaluation," International Journal of Information Technology & Decision Making, vol. 17, no. 2, pp. 513–525, 2018, doi: 10.1142/S0219622017500237.

[36] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: case study in Malaysia," Developments in the Built Environment, vol. 5, p. 100037, 2021, doi: 10.1016/j.dibe.2020.100037.

[37] M. H. L. Lee, Y. C. Ser, G. Selvachandran, P. H. Thong, L. Cuong, L. H. Son, N. T. Tuan, and V. C. Gerogiannis, "A comparative study of forecasting electricity consumption using machine learning models," Mathematics, vol. 10, no. 8, p. 1329, 2022, doi: 10.3390/math10081329.

[38] World Integrated Trade Solution, "Cereals; husked (brown) rice exports by country in 2019," *World Bank*, May 7, 2024. [Online]. Available: https://wits.worldbank.org/trade/comtrade/en/country/ALL/year/2019/tradeflow/Exports/partner/WLD/product/100620#:~:text=In%202019%2C%20Top%20exporters%20of,99%2C580.45K%20%2C%20126%2C817%2C000%20Kg. [Accessed: Nov. 13, 2024].

[39] International Rice Research Institute, "IRRI commends Indonesia for strong efforts to achieve rice self-sufficiency," *CGIAR*, Aug. 16, 2022. [Online]. Available: https://www.cgiar.org/news-events/news/irri-commends-indonesia-for-strong-efforts-to-achieve-rice-self-sufficiency/. [Accessed: Nov. 13, 2024].

[40] S. Aronhime *et al.*, "DCE-MRI of the liver: effect of linear and nonlinear conversions on hepatic perfusion quantification and reproducibility," Journal of Magnetic Resonance Imaging, vol. 40, no. 1, pp. 90–98, 2014, doi: 10.1002/jmri.24341.

[41] United States Department of Agriculture, "Laos rice report annual," 2020. [Online]. Available: https://apps.fas.usda.gov/newgainapi/api/Report/DownloadReportByFileName?fileName=Laos%20Rice%20Report%20Annual_Bangkok_Laos_06-08-2020#:~:text=Report%20Highlights%3A,to%201.5%20million%20metric%20tons. [Accessed: Nov. 13, 2024].

[42] H. Chanthou, "Is Cambodian rice ready for the world market?" [Online]. Available: https://www.adb.org/multimedia/partnership-report2021/stories/is-cambodian-rice-ready-for-the-world-market/. [Accessed: Nov. 13, 2024].

[43] "2025 target for Cambodia to export 1 million tonnes of rice," *Khmer Times*, Feb. 12, 2023. [Online]. Available: https://www.khmertimeskh.com/501237171/2025-target-for-cambodia-to-export-1-million-tonnes-of-rice/. [Accessed: Nov. 13, 2024].

[44] Department of Statistics Malaysia, "Selected agricultural indicators, Malaysia, 2021," Oct. 25, 2022. [Online]. Available: https://www.dosm.gov.my/portal-main/release-content/selected-agricultural-indicators-malaysia-2021. [Accessed: Nov. 13, 2024].

[45] Department of Statistics Malaysia, "Selected agricultural indicators, Malaysia, 2023," Oct. 27, 2023. [Online]. Available: https://www.dosm.gov.my/portal-main/release-content/selected-agricultural-indicators-malaysia. [Accessed: Nov. 13, 2024].

[46] S. Kasinathan, "National audit report shows nearly quarter of Malaysian paddy farmers earn below RM600 monthly, as rice cultivation programme fails to reach target," *Malay Mail*, Nov. 22, 2023. [Online]. Available: https://www.malaymail.com/news/malaysia/2023/11/22/national-audit-report-shows-nearly-quarter-of-malaysian-paddy-farmers-earn-below-rm600-monthly-as-rice-cultivation-programme-fails-to-reach-target/103487. [Accessed: Nov. 13, 2024].