Journal of Quality Measurement and Analysis *JQMA* 21(4) 2025, 55-76
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*
https://doi.org/10.17576/jqma.2104.2025.04

ENHANCING PREDICTIVE PERFORMANCE IN STATISTICAL MODELING: INNOVATIVE HYBRID BEST SUBSET FEATURE SELECTION FOR RICE PRODUCTION IN MALAYSIA

(Mempertingkatkan Prestasi Ramalan dalam Pemodelan Statistik: Pemilihan Fitur Hibrid Subset Terbaik Inovatif bagi Pengeluaran Beras Di Malaysia)

ZUN LIANG CHUAN*, ABRAHAM LIM BING SERN, REN SHENG THAM, DAVID LAU KING LUEN & CHEK CHENG TAN

ABSTRACT

Recent statistics from the Food and Agriculture Organization (FAO) of the United Nations revealed an increasing trend in both moderate and severe food insecurity prevalence in Malaysia on average. To effectively address this issue, comprehensive solutions are needed to consider the four dimensions of food security, particularly for predicting rice production, a staple food for Malaysians. This study aimed to propose an innovative hybrid deterministic best subset feature selection method for identifying significant determinants impacting rice production in Malaysia, thereby contributing to a more effective understanding and management of food security. These selected determinants aligned with the four dimensions of food security and the key pillars of the Sustainable Development Goals (SDGs). The proposed feature selection method integrates mathematics techniques, specifically the modified Taguchi-based VIseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) multi-criteria decision-making (MCDM) algorithm and three performance metrics. The analysis demonstrated that the proposed method outperformed existing hybrid deterministic feature selection methods in the literature, which lacked a comprehensive consideration of all dimensions of food security. Furthermore, the analysis revealed that the proposed methods consistently achieved higher accuracy compared to automated deterministic wrapper feature selection methods. This article has principally contributed to both the mathematics academia and industry realms. This study provided valuable insight for academicians, policymakers, smallholder farmers, and society by offering a more effective feature selection method, thereby enhancing policy development and farming practices in the context of food security. It contributed significantly to both academic and industry realms by presenting a hybrid deterministic features selection method that enhanced communication and practical application compared to the stochastic metaheuristic features selection method.

Keywords: food security; rice production; statistical modeling; feature selection; predictive performance

ABSTRAK

Statistik kebelakangan ini dari Organisasi Makanan dan Pertanian (FAO) Persatuan Bangsa-Bangsa Bersatu memaparkan rentetan peningkatan terhadap prevalens ketidakselamatan makanan secara purata bagi kedua-dua tahap sederhana dan parah di Malaysia. Bagi menangani isu ini secara berkesan, penyelesaian menyeluruh amat diperlukan dengan mengambil kira keempat-empat dimensi keselamatan makanan, khususnya peramalan pengeluaran beras, yang merupakan makanan ruji dalam kalangan rakyat Malaysia. Kajian ini bermatlamat mengusulkan suatu kaedah pemilihan fitur hibrid berketentuan subset terbaik inovatif bagi mengenal pasti penentu-penentu signifikan yang mengimplikasikan pengeluaran beras di Malaysia, dengan itu menyumbang kepada pemahaman dan pengurusan keselamatan makanan dengan lebih berkesan. Penentu-penentu terpilih adalah tekal dengan keempat-empat dimensi keselamatan makanan dan tetiang Matlamat Pembangunan Mampan (SDGs). Kaedah pemilihan fitur yang diusulkan bersepadu teknik matematik, khususnya algoritma pembuat keputusan berbilang kriteria *VIseKriterijumska Optimizacija I Kompromisno Resenje*

(VIKOR) berasaskan Taguchi terubah suai, dan tiga metrik prestasi. Hasil analisis memaparkan kaedah yang diusulkan mengatasi kaedah pemilihan fitur hibrid berketentuan dalam sorotan kajian, yang tidak mengambil kira keempat-empat dimensi keselamatan makanan secara menyeluruh. Tambahan pula, hasil analisis mendedahkan kaedah pemilihan fitur hibrid berketentuan yang diusulkan secara tekal mecapai ketepatan yang lebih tinggi berbanding kaedah pemilihan fitur pembungkusan automatik. Makalah ini memberi sumbangan dalam kedua-dua bidang akademik berteraskan matematik dan industri. Kajian ini menyediakan pencerahan yang berharga kepada ahli akademik, penggubal dasar, petani kecil, dan masyarakat dengan menawarkan suatu kaedah pemilihan fitur yang berketerampilan, dengan mempertingkatkan pembangunan dasar dan amalan pertanian dalam konteks keselamatan makanan. Ia juga menyumbangkan secara signifikan dalam kedua-dua bidang akademik dan industri dengan melakarkan kaedah pemilihan fitur berketentuan yang mempertingkatkan komunikasi dan aplikasi gunaan dibandingkan dengan kaedah pemilihan fitur metaheuristik stokastik.

Kata kunci: keselamatan makanan; pengeluaran beras; pemodelan statistik; pemilihan fitur; prestasi ramalan

1. Introduction

Statistics from the Food and Agriculture Organization (FAO) of the United Nations (Chuan *et al.* 2025; Food and Agriculture Organization of the United Nations 2024) reveal an increasing trend in both moderate and severe food insecurity in Malaysia from 2015 to 2023. The 3-year average prevalence in the total population (%) increased by 1.6%, rising from 15.1% in 2016 (average spanning 2015-2017) to 16.7% in 2022 (average spanning 2021-2023). Although these figures are slightly lower than the 3-year average prevalence in 2015 (17.4% with an average spanning 2014-2016), they signaled an early alarm that required attention from academicians, policymakers, smallholder farmers, researchers, and society. Addressing food insecurity is crucial for national economic growth and sustainability, aligning with the Madani Economy Framework (Wong *et al.* 2024) and the Twelfth Malaysia Plan 2021-2025 (12MP), which aimed to revitalize the agriculture sector, support downstream activity, and enhance food security (Economic Planning Unit Prime Minister's Department 2021; Ministry of Economy 2023; Wong *et al.* 2024), particularly focusing on staple foods such as rice production.

Given the rising food insecurity trends, a comprehensive approach is needed to tackle these challenges, involving a deeper understanding of the multidimensional aspects of food security. The FAO defined four key dimensions of food security: availability, accessibility, stability, and utilization, aligning with the four key pillars of the Sustainable Development Goals (SDGs). Availability refers to sufficient quantities of appropriate food available consistently from domestic production, commercial imports, food assistance, or food services. Accessibility indicates that people have adequate incomes or other resources to access appropriate food domestically through home production, purchasing in local markets, exchanging gifts, borrowing, or as food aid. Stability ensures that food is adequately available all the time, so access and utilization are not curtailed by shortages, emergencies, or crises. Utilization involves proper storage and processing of food, with sufficient knowledge applied nutritionally, health-wise, and according to social-cultural and spiritual considerations (Chuan *et al.* 2025; Gunaratne *et al.* 2021).

Understanding these dimensions highlighted the necessity for a comprehensive predictive model for rice production that integrates a broad range of determinants to address food

security comprehensively. This emphasizes the importance of feature engineering, such as feature selection in predictive modeling. Effective feature selection is crucial to managing information redundancy, optimizing computational resources, enhancing predictive performance, and simplifying interpretation and communication from a practical perspective. However, the literature on feature selection specifically for rice production prediction is limited. Previous studies have primarily utilized the intuitive feature selection methods, best subset feature selection method with physical interpretation, filter feature selection methods, and hybrid wrapper-filter feature selection methods. These include hybrid automated forward feature selection combined with Student's *t*-test (A7), hybrid automated backward elimination feature selection with Student's *t*-test (A8), and automated stepwise feature selection with Student's *t*-test (A9). Despite these existing methods, there remains a significant research gap in applying more automated and sophisticated feature selection methods, such as the best subset wrapper feature selection method, in agricultural economics and engineering studies.

To address this gap, this article proposed an innovative hybrid deterministic feature selection method combining automated best subset and Student's *t*-test. The automated best subset feature selection is based on performance metrics like the coefficient of determination (Adj-R^2), Mallow's Cp (C_p), and Bayesian Information Criterion (BIC), followed by filtering insignificant determinants utilizing Student's *t*-test. To validate this feature selection method, this study benchmarked it against previous feature selection approaches and evaluated the forecasting accuracy utilizing an Ordinary Least Squares Multiple Linear Regression (OLS-MLR) predictive model. The effectiveness of these methods is demonstrated through the modified Taguchi-based VIseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) multi-criteria decision-making (MCDM) algorithm.

While various feature selection methods have been explored, this study focused on deterministic methods, contrasting them with statistical and sophisticated computing feature selection methods. Features selection methods such as Correlation-Based Feature Selection (CBFS), Variance Inflation Factor (VIF), Random Forest Variance Important (RFVarImp) (Maya Gopal & Bhargavi 2019), multi-stage hybrid feature selection methods (Mishra *et al.* 2021), ensemble feature selection methods (Sathya & Gnanasekaran 2023), Artificial Bee Colony (ABC) algorithm-based feature selection method (Manjunath & Pallayan 2024), and correlation analysis (Wijayanti *et al.* 2024) relied on stochastic mechanisms and are widely proposed in the literature. However, these methods are deemed inappropriate due to their stochastic nature, lack of statistical evidence for selected determinants, and complexity in interpreting and communicating the selected determinants for real-life applications, making them impractical for feature selection. Hence, deterministic feature selection methods are favored for reliability and clarity.

In summary, the deterministic feature selection method proposed in this study significantly contributes to both the mathematical academic field and the agricultural economic and engineering industry. Specifically, for the mathematical field, this study introduces an innovative hybrid deterministic feature selection approach that enhances the development of parsimonious predictive models, which are valued for their simplicity and effectiveness. Moreover, by addressing a notable gap in the literature, this study demonstrates practical applications that provide valuable insights for various stakeholders, including academicians, policymakers, smallholder farmers, researchers, and society. These insights are crucial for enhancing policy development and improving farming practices in the context of food security. The remainder of this article is organized as follows. Section 2 provides an overview of literature reviews, Section 3 outlines the research methodology and theoretical background, Section 4 presents the analysis results, and Section 5 offers concluding remarks.

2. Related Works

Feature selection methods in statistical learning are categorized into three principal approaches: filter, wrapper, and embedded. The filter feature selection approach assesses determinant sets independently of statistical learning predictive models or algorithms, utilizing statistical metrics such as correlation analysis, and parametric and non-parametric statistical hypothesis testing to evaluate the association between determinants and the endogenous variable. The wrapper feature selection approach, on the other hand, employs a greedy search to evaluate all possible combinations of determinants based on specific evaluation metrics, involving training and evaluating a predictive model for each subset of determinants. The embedded feature selection approach integrates feature selection within statistical learning predictive models or algorithms, such as the Least Absolute Shrinkage and Selection Operator (LASSO) for linear regression and decision trees that assess determinant importance during training.

Despite their significance and widespread utilization in various domains, feature selection methods have received limited attention in specific fields such as agriculture economics and engineering, particularly in the context of predicting rice production. Studies in these areas have predominantly utilized intuitive feature selection methods, best subset feature selection with physical interpretation, filter feature selection methods, and hybrid wrapper-filter feature selection methods. For instance, intuitive feature selection methods involve selecting determinants based on their apparent relevance to endogenous variables such as paddy and rice production and productivity, guided by insights from existing literature rather than formal algorithms or statistical testing.

To further address this gap, it is notable that most of the Malaysian literature focuses on univariate paddy and rice production predictions utilizing conventional statistical time-series predictive models such as Autoregressive Integrated Moving Average (ARIMA) (Ahmad *et al.* 2017; Fauzi & Bakar 2022) and Naïve approach (Yusof *et al.* 2019), frequently without incorporating feature engineering. For instance, Ahmad *et al.* (2017) proposed utilizing the ARIMA (0,2,2) model to predict annual paddy production in the Muda Agricultural Development Authority (MADA) region from 1979 to 2014. While their research provided valuable insights for refining paddy production policies, the univariate analysis was limited as it did not consider various determinants impacting paddy production. Additionally, the article presented a misconception by assuming that estimated values were equivalent to forecasted values in some parts and that the utilization of a small sample size (<50) dataset could lead to biased results when employing maximum likelihood estimation (MLE).

Similarly, Fauzi and Bakar (2022) proposed utilizing the ARIMA (1,1,0) model to predict annual rice production in Malaysia from 1980 to 2021. This study, which also involved a small sample size dataset, indicated that the predictive results were primarily beneficial for policymakers, researchers, and rice farmers by providing insights for short-run planning and budgeting. However, this study shared similar limitations with Ahmad *et al.*'s research, particularly its reliance on univariate analysis. Moreover, neither study adequately discussed diagnostic checking of the predictive model. Notably, Fauzi and Bakar's forecasted trends showed a consistently decreasing straight trend line, suggesting that the models might have been inadequate due to their failure to capture the variability and complexity of rice production.

Conversely, Yusof *et al.* (2019) employed the Naïve approach to forecast paddy productivity in the Kedah rice bowl of Malaysia. They aggregated the monthly dataset into annual totals for the period from 2005 to 2012, distinguishing between the main season (August-February) and off-season (March-July), and forecasted paddy productivity for the

subsequent four years. Notably, this study did not validate its results utilizing other univariate statistical time-series predictive models. This study concluded that precipitation does not directly impact paddy productivity, which may lead to biased and subjective conclusions. The analysis was primarily based on comparisons between the main and off-seasons without accounting for atmospheric conditions such as precipitation, offering merely a limited and subjective perspective. Additionally, this univariate approach shares similar limitations with the studies by Ahmad *et al.* (2017), and Fauzi and Bakar (2022) studies, providing limited insights into food security.

In contrast, various studies have employed econometrics predictive models (Alam *et al.* 2011; Chuan *et al.* 2022; Makhtar *et al.* 2022; Tan *et al.* 2021), statistical multivariable predictive models (Alam *et al.* 2014; Idalisa *et al.* 2019; Putri *et al.* 2019; Shahidan *et al.* 2022), and statistical multivariate predictive models (Roslan *et al.* 2021) to address the principal limitations of univariate analysis. Alam *et al.* (2011) proposed an econometric predictive model to analyze the association between paddy production and various determinants related to atmospheric conditions, socioeconomic factors, and farming practices, focusing on North-West Selangor. Their approach utilized a cross-sectional OLS-MLR predictive model. However, this model based on intuitive feature selection, may lead to sub-optimal analysis results. Additionally, this study relied solely on survey data collection rather than incorporating time-series analysis, limiting the model's ability to fully address the data complexities. Chuan *et al.* (2022) employed a more comprehensive approach by modeling a cross-sectional time-series dataset. They utilized a one-way random effects econometric panel regression analysis to investigate the agricultural economic activities of C3 plants, such as cocoa, natural rubber, oil palm, and rice. This study integrated climatic and non-climatic determinants, applying a filter feature selection method. Despite this sophisticated approach, the coefficient of determination (not determined) indicated that the predictive model had limited predictive capability.

Tan *et al.* (2021) proposed a one-way fixed-effect panel analysis to regress atmospheric conditions (minimum temperature, maximum temperature, and precipitation) and socioeconomic factors (total cultivated area) determinants against rice productivity, focusing on eight granary areas in Peninsular Malaysia. The analysis results indicated that the dataset was well-fitted for the proposed fixed-effect panel analysis, as evidenced by a high Adj-R^2 exceeding 90% for the main season. However, this study offered limited insights into effectively addressing food security, as it merely covered two of the four key dimensions of food security: availability and stability. Additionally, the limited number of atmospheric conditions and socioeconomic factors determinants considered may lead to sub-optimal results, especially if decision-making overly relies on the coefficient of determination. Furthermore, the proposed predictive model fails to account for the adjustment effects of other determinants, including those related to farming practices.

The multivariate econometrics predictive model, such as the Vector Error Correction Model (VECM) has also been employed to predict rice productivity by calibrating total rice production per hectare rather than overall rice production. Specifically, Makhtar *et al*. (2022) utilized VECM to identify key determinants impacting rice productivity, including annual average farmer income, Food Consumer Price Index (CPI), granary area, labor in agriculture, forestry, and fishing, paddy yield, rice import dependency, and rice production. They applied the natural logarithm ($\log(\bullet)$) function to each original determinant, and the feature selection method in this study was based on an intuitive approach. The analysis results identified the paddy granary area, paddy yield, rice production, and CPI as relevant to rice productivity in the short-run. However, this approach may introduce bias. This article assumes that

multivariate analysis is equivalent to multivariable analysis, which can lead to misconceptions. This is particularly evident as this study presents merely one Error Correction Model (ECM) and omits the key determinants of rice production, despite considering seven original determinants. This may reflect misunderstandings from both econometrics and statistical perspectives. In simple terms, the multivariate analysis did not include determinants, unlike a true multivariable analysis.

On the other hand, Alam *et al.* (2014) employed OLS-MLR and Ordinary Least Squares Multiple Nonlinear Regression (OLS-MNLR) predictive models to investigate the association between average paddy yield per season and atmospheric conditions, such as average monthly rainfall and daily temperature, as well as technology utilization, focusing on North West Selangor. This study utilized an intuitive feature selection method, and their analysis indicated that temperature and technology had statistically significant determinantal and favorable impacts on paddy production according to the OLS-MLR predictive model. Similarly, all three determinants showed statistically significant determinantal and favorable impacts according to the OLS-MNLR predictive model. However, this study did not conclude which predictive model was superior. Additionally, the reliance on microdata, the intuitive feature selection method, and the failure to conduct diagnostic checks may have introduced bias into the analysis, leading to limited insights regarding food insecurity.

Meanwhile, Idalisa *et al.* (2019) also employed the OLS-MLR predictive model to analyze the association of rice production with determinants such as paddy production, planted area, human population, and domestic consumption, utilizing an intuitive feature selection method. The central focus of their study was on the parameter estimation method for the OLS-MLR predictive model. Specifically, they proposed estimating the OLS-MLR predictive model's parameters utilizing two variants of the Conjugate Gradient (CG) numerical method – Fletcher and Reeves (FR), and Polak, Ribiere, and Polyak (PRP)) – with the Ordinary Least Squares (OLS) estimation method serving as a benchmark. However, relying on both the intuitive feature selection method and the CG numerical method raises concerns. While intuitive feature selection can be useful in certain situations, it may not frequently capture the underlying probabilistic associations within the datasets, potentially leading to incomplete predictive models and impacting the accuracy of parameter estimates. Additionally, utilizing the CG numerical method to estimate parameters in a probabilistic predictive model might result in estimates that do not meet the Minimum Variance Unbiased Estimator (MVUE) criteria, which is fundamental in statistical theory. This could result in parameter estimates that are not merely biased but also exhibit greater variance than theoretically optimal, potentially compromising the model's effectiveness.

Putri *et al.* (2019) conducted a regression analysis investigating the association of rice production with crop health indicators – such as Soil Plant Analysis Development (SPAD) readings at 45, 70, and 90 days after planting – and soil nutrients, including total nitrogen content (%), available phosphorus (mg/kg), and exchangeable potassium (cmol/kg). This study utilized statistical OLS-MLR predictive models based on a dataset collected from Sungai Besar, Selangor. Five different OLS-MLR predictive models were developed utilizing both intuitive and correlation analysis filter feature selection methods. The results indicated that the predictive model developed utilizing the intuitive feature selection method outperformed the predictive model based on the correlation filter. However, the absence of statistical evidence for selected determinants, the consideration of interactions between determinants, and the omission of diagnostic checks to validate OLS assumptions could present challenges in interpreting and applying the findings in practical contexts.

In a similar vein, Shahidan *et al.* (2022) proposed regressing and forecasting paddy production by utilizing the rice harvested area, year, and cropping intensity factor (%) for Melaka from 1988 to 2030, utilizing the OLS-MLR predictive model. They computed rice production through a mathematical deterministic conversion equation, where rice production is derived by multiplying paddy production by the paddy-to-rice conversion rate. The determinants included in their OLS-MLR predictive model were selected based on an intuitive approach. Although this study incorporated necessary diagnostic checks, the feature selection method utilized could lead to biased outcomes. Additionally, this study did not account for potential determinants relevant to the four dimensions of food security, which limits its ability to provide a comprehensive analysis of the determinants significantly impacting paddy and rice production.

Nevertheless, statistical multivariate analysis has also been employed in modeling and forecasting paddy production for five selected Southeast Asia nations (Myanmar, Vietnam, Indonesia, Malaysia, and Thailand). Notably, Roslan *et al.* (2021) utilized two variants of the multivariate probability distribution function: the Elliptical Copula family (Normal and *t*-Copula), and the Archimedean Copula family (Joe, Clayton, and Gumbel) to identify key determinants from 1961 to 2013. OLS-MLR and multivariate normal distribution (MVN) predictive models served as benchmarks. The original determinants considered included paddy production, planted area, fertilizer usage, total annual average rainfall, and maximum average temperature. Ten univariate statistical distributions were employed to fit each original determinant, with the distribution minimizing the Akaike Information Criterion (AIC) selected for copula modeling. This study utilized the best subset feature selection method with physical interpretation, investigating 15 possible combinations of the original determinants, with rice production included in each predictive model. Despite the high capability of Copula modeling in predicting paddy production across the selected lower-middle and upper-middle Southeast Asia nations compared to OLS-MLR and MVN predictive models, this study also failed to account for potential determinants encompassing all four dimensions of food security, resulting in limited insight into addressing food security.

Recently, AI-based predictive algorithms for rice production have garnered attention from researchers in Malaysia. For instance, Marong *et al.* (2024) compared several multivariable AI-based predictive algorithms, including Support Vector Regression (SVR), Random Forest (RF), and Artificial Neural Networks (ANN), for predicting rice production based on atmospheric conditions determinants. Their study utilized an intuitive feature selection method to select atmospheric conditions determinants (rainfall, temperature, humidity, and flood datasets) for the proposed AI-based predictive algorithms, except for RF, which utilized automated feature selection. Although RF was found to outperform SVR and ANN predictive algorithms, this study noted that RF's automated feature selection might lead to biased results due to its stochastic nature. Furthermore, the automated feature selection lacked statistical validation for selecting determinants and faced limitations in forecasting both short-run and long-run rice production. This article also misinterpreted forecasting accuracy evaluations and presented incorrect mathematical equations, particularly for Goodness-of-Fit (GoF) measures. Additionally, while the article claimed to propose a hybrid AI-based predictive algorithm combining machine learning (ML) and deep learning (DL), it did not provide performance results for this hybrid approach.

Marong *et al.* (2024) asserted that the OLS-MLR predictive algorithm failed to capture complex nonlinear relationships in rice production without effectively comparing it to Malaysia's dataset. This claim was not supported by existing literature. Specifically, Chuan *et al.* (2024c) compared OLS-MLR and OLS-MNLR predictive algorithms for rice production.

They investigated the association between rice production and atmospheric conditions determinants (annual precipitation, annual maximum temperature, and Carbon Dioxide (CO_2) emission) as well as non-atmospheric conditions determinants (annual population, planted area, and gross domestic product (GDP)) utilizing OLS-MLR and OLS-MNLR predictive algorithms, respectively. To ensure the inclusion of parsimonious and significant determinants, they compared hybrid automated wrapper feature selection methods (forward feature selection (FFS), backward feature selection (BFS), and stepwise feature selection (SFS)), with the filter feature selection method (Student's *t*-test). Their analysis found that OLS-MLR predictive algorithms combined with BFS and SFS, along with the Student's *t*-test outperformed other AI-based predictive algorithms. They concluded that atmospheric conditions determinants did not significantly impact rice production in Malaysia, while non-atmospheric conditions determinants, such as planted area and annual population had a significant impact. This study highlighted that OLS-MLR is suitable for predicting rice production in Malaysia.

Building on Chuan *et al.*'s work, Chuan *et al.* (2025) proposed regressing the association of rice production with various determinants: atmospheric (annual mean maximum temperature, average annual precipitation, and CO_2 emissions), socioeconomic (crude oil price, domestic supply, food supply, import, inflation consumer prices, labor index, population, and urbanization rate), and farming practices (agriculture land index, land use, and machinery per agriculture land) across nine lower-middle and upper-middle Southeast Asia nations utilizing a modified stacked ensemble MLR-SVR-based predictive algorithm was noted for its robustness to the presence of outliers. In this study, they considered three dimensions of food security: availability, accessibility, and stability. Given the medium-dimensional nature of their dataset, they employed the hybrid wrapper-filter feature selection methods. Their analysis revealed that the determinants impacting rice production across these Southeast Asian nations are not uniform, largely due to geographical, economic, and agricultural policy factors.

Due to the effective analysis presented in Chuan *et al.*'s works, this study aims to extend their research by incorporating an additional dimension of food security, specifically the utilization dimension, into feature selection. The inclusion of four dimensions of food security has the potential to involve numerous determinant sets, which may introduce redundancy and include determinants that are insignificant to the rice production predictive model. To address this issue, this study proposed applying a hybrid automated best subset wrapper features selection method combined with the Student *t*-test, a technique that has not been previously explored or discussed in agricultural economic and engineering studies. Unlike the previous study that focused on AI-based predictive algorithms, this study emphasizes statistical modeling and, therefore, does not implement a training and test set split. This presents a challenge for direct comparisons between statistical and AI-based predictive algorithms. The rationale behind utilizing statistical modeling is to ensure that the selected determinants are also applicable in a Bayesian OLS-MLR predictive model.

Nevertheless, various statistical and sophisticated computing feature selection methods for rice production and productivity prediction have been proposed in the international literature. These include multi-stage hybrid feature selection methods such as Correlation-Based Feature Selection (CBFS) feature selection, Variance Inflation Factor (VIF) feature selection, Random Forest Variance Important (RFVarImp) feature selection, Borda count-based feature ranking, and feature fusion strategy, ensemble feature selection methods, ABC algorithm-based feature selection method, and correlation analysis. However, these sophisticated feature selection methods proposed by computer science applications are frequently impractical for real-life

applications due to the involvement of subjective interpretation and decision-making, complexity, and stochastic nature (except CBFS, VIF feature selection, and correlation analysis), which could lead to biased analysis and insights into agricultural economics and engineering. Moreover, ensemble feature selection methods associated with machine learning classifiers are inadequately utilized in fitting time-series datasets due to the loss of temporal insights. Consequently, this study does not consider these sophisticated feature selection methods as benchmarks, opting instead for deterministic feature selection methods that are more appropriate for this research focus.

3. Research Methodology

This study employed the Cross Industry Standard Process for Data Mining (CRISP-DM) framework, which comprised six principal phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Since the key contribution of this study focused on feature selection within the data preparation phase, the theoretical background of the proposed innovative hybrid deterministic wrapper-filter feature selection method was detailed in Sections 3.1–3.2. Section 3.3 provided a schematic overview of the proposed innovative hybrid feature selection method, aligned with the CRISP-DM framework, with a focus on identifying a parsimonious determinant set.

CRISP-DM was selected over alternative frameworks, such as Knowledge Discovery in Databases (KDD), and Sample, Explore, Modify, Model, and Assess (SEMMA), due to its iterative flexibility and comprehensive nature. The CRISP-DM framework's success has spanned a range of research fields, including computer science (Garcia-Arteaga *et al.* 2024), energy economics (Chuan *et al.* 2024b), education economics (Chuan *et al.* 2024a; Liang *et al.* 2024), finance research (Cheng 2023), forensic science (Chuan *et al.* 2023), and healthcare research (Lohaj *et al.* 2023).

3.1. *Theoretical background of the best subset feature selection method*

The OLS-MLR is a widely utilized model for predicting an endogenous variable based on a set of independent determinants in limited agricultural economics and engineering literature. Suppose that $\mathbf{Y}_{\text{obs}} = \left[y_i \right]_{n \times 1}; i = 1, 2, \ldots, n$ represents a vector of rice production sample dataset of size n, and $\mathbf{X} = \left[\mathbf{1X}_j \right]_{n \times (k+1)}; \mathbf{X}_j = \left[x_{ij} \right]_{n \times 1}, j = 1, 2, \ldots, k$ represents a matrix comprised k determinants with each jth determinants of size n. Therefore, the algebraic form of the population OLS-MLR predictive model can be expressed as Eq. (1).

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{\beta} = \left[\beta_j \right]_{(k+1) \times 1}$ represents the vector of OLS-MLR predictive model parameters. In achieving the Best Linear Unbiased Estimation (BLUE) properties, $\boldsymbol{\beta}$ is estimated by utilizing the OLS estimation method, such that

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{Y} \tag{2}$$

which is guaranteed by the Gauss-Markov Theorem (GMT). In simple terms, the assumptions of errors, such that $\boldsymbol{\varepsilon} \xrightarrow{i.i.d} N\left(0,\sigma^2\right)$ should be satisfied to yield a reliable OLS-MLR predictive model, where $i.i.d$ represents the independent and identically distributed.

However, including all potential determinants in an OLS-MLR predictive model can lead to high computational costs and inefficiency, violating the principle of parsimony, which seeks a simple and effective predictive model. This could undermine the predictive model's reliability and fail to meet the assumptions of the GMT. To overcome this, a deterministic feature selection method is essential for identifying the most significant determinants, creating a parsimonious, computationally efficient, and interpretable predictive model.

Suppose that $\mathbf{X}_p = \left[\mathbf{1}\mathbf{X}_{j'}\right]_{n\times(p+1)}$ represents a reduced matrix consisting of p determinants, where $p \le k$. $\mathbf{X}_p$ is derived utilizing a wrapper feature selection method, such as the best subset approach. In statistics theory, the best subset of $\mathbf{X}_p$ is selected based on all possible combinations of p determinants, and evaluated utilizing three well-known performance metrics: Adj-R^2, C_p, and BIC. These metrics are mathematically defined as Eqs. (3)–(5).

$$\text{Adj}-\text{R}^2 = 1 - \frac{ESS_p\left(n-1\right)}{TSS_p\left(n-p-1\right)} \tag{3}$$

$$\text{C}_\text{p} = \frac{RSS_p}{\hat{\sigma}_k} + \left(2p-n\right) \tag{4}$$

$$\text{BIC} = -2L_p + \left(p+1\right)\log\left(n\right) \tag{5}$$

where ESS_p is the error sum of squares of the reduced predictive model of OLS-MLR, which comprised p determinants, TSS_p is the total sum of squares of the reduced predictive model, RSS_p is the regression sum of squares, $\hat{\sigma}_k^2$ is the error mean of squares for the full predictive model, which comprised k determinants, $L_p = -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\hat{\sigma}_p^2\right) - \frac{1}{2\hat{\sigma}_p^2}RSS_p$ is the log-likelihood function value for the reduced predictive model, and $\hat{\sigma}_p^2$ is the error mean of squares for the reduced predictive model.

In statistical modeling, an Adj-R^2 value approaching 1 indicates a well-fitted OLS-MLR predictive model. A higher Adj-R^2 also suggests that a particular combination of determinants forms the best subset compared to other potential subsets. On the other hand, lower values of C_p and BIC indicate the best subset compared to other potential subsets. While a low C_p value aligns with the principle of parsimony, it is also essential for C_p to be close to p, as this minimizes predictive model bias and improves its reliability.

3.2. *Theoretical background of the Student's t-test feature selection method*

While the best subset of determinants identified in Section 3.1 is useful for reducing the determinants' space, it does not necessarily imply that each determinant in the best subset is individually significant to the OLS-MLR predictive model. To address this limitation, this article proposes the utilization of the Student's *t*-test to evaluate the individual statistical significance of each determinant in the reduced predictive model, resulting in a more parsimonious predictive model that adheres to the principle of parsimony. From a theoretical

perspective, the Student's *t*-test evaluates the statistical significance of each jth determinant by comparing the estimated parameter to its standard error utilizing Eq. (6).

$$\text{Student's } t\text{ - test} = \frac{\hat{\beta}_j}{\sqrt{\left(\frac{RSS_p}{n-p-1}\right)\left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1}}} \tag{6}$$

where $\hat{\beta}_j$ represents the OLS-MLR parameter estimate for the jth determinant in the reduced predictive model. The Student's *t*-test is preferred over the *F*-test for feature selection because the *F*-test evaluates the overall significance of the determinant set but does not assess the significance of individual determinants. The determinants identified as statistically significant through the Student's *t*-test are retained in the reduced predictive model, which is then subjected to diagnostic checks and internal forecasting accuracy evaluations. This process ensures that the final predictive model is both effective and parsimonious, aligning with the principle of parsimony.

3.3. *The procedure of the proposed innovative hybrid feature selection method*

This section outlines the step-by-step procedure for implementing the proposed innovative hybrid deterministic best subset and Student's *t*-test feature selection method:

Step 1: Acquire relevant open-source datasets covering the period 1961-2021 from databases such as the Climate Change Knowledge Portal (CCKP), Our World in Data (OWID), and the World Bank, as outlined in Table 1.

Step 2: Compile and consolidate the datasets acquired from multiple sources into a single Comma-Separated Values (CSV) file to ensure consistency and facilitate analysis.

Step 3: Perform a preliminary statistical analysis to assess dataset quality and gain statistical insights. This includes tabulation summarization, visualizing correlations utilizing correlograms, conducting correlation analysis, and computing the first four L-moments (Hosking 1990; Chuan *et al.* 2024b, 2025) for further dataset characterization.

Step 4: Detect outliers utilizing the Interquartile Range (IQR) method, where the 1.5IQR rule identifies mild outliers, and the 3IQR rule detects extreme outliers.

Step 5: Apply the capping method to correct extreme outliers by replacing values beyond the 3IQR boundaries with the respective lower and upper outer fences, if necessary.

Step 6: Select an optimal subset of determinants utilizing the best subset wrapper feature selection method, with Adj-R^2 as the performance metric.

Step 7: Refine the selected optimal determinant set by eliminating statistically insignificant determinants through the Student's *t*-test filter feature selection method, ensuring that merely relevant determinants remain.

Step 8: Train an OLS-MLR predictive model utilizing the final set of selected determinants.

Step 9: Assess the validity of the trained OLS-MLR predictive model by performing diagnostic tests to ensure key statistical assumptions are met. Utilize the Shapiro-Wilk test to check if the residuals follow a normal distribution. Apply the Breusch-Pagan test to verify homoscedasticity, ensuring that residual variance remains stable. Conduct the Run test to verify the independence of residuals and rule out

autocorrelation. Finally, compute the Variance Inflation Factor (VIF) to detect multicollinearity, utilizing a threshold of exceeding 10 as a criterion of concern.

Table 1: Classification of key determinants in relation to SDGs pillars and food security dimensions

SDGs's Pillar	Variable (Measurement unit)	Notation	Food Security Dimensions
Environmental	Annual CO_2 emissions (per capita)	x_1	Availability and Stability
	Average Annual Maximum Temperature (˚C)	x_2	Availability and Stability
	Average Annual Minimum Temperature (˚C)	x_3	Availability and Stability
	Average Annual Precipitation (millimeters)	x_4	Availability and Stability
	Fertilizer utilizes: Nutrient Nitrogen (kilograms per hectare)	x_5	Availability
	Fertilizer utilizes: Nutrient Phosphate (kilograms per hectare)	x_6	Availability
	Fertilizer utilizes: Nutrient Potash (kilograms per hectare)	x_7	Availability
	Land use (per capita)	x_8	Availability
Economic	Annual Growth of GDP (per capita)	x_9	Availability
	Area harvested (hectare per capita)	x_{10}	Availability
	Consumer Price Index	x_{11}	Accessibility
	Gross National Income Growth (Annual %)	x_{12}	Accessibility
	Import of goods and services (constant 2015 US$)	x_{13}	Accessibility
	Inflation in Consumer Price (annual %)	x_{14}	Accessibility and Stability
	Trade (% of GDP)	x_{15}	Availability
	World crude oil prices (US$)	x_{16}	Accessibility and Stability
Human	Cereals allocated to human food (billion tones)	x_{17}	Utilization
	Daily per capita calorie supply (kilocalories)	x_{18}	Accessibility and Utilization
	Daily per capita fat supply (grams)	x_{19}	Accessibility and Utilization
	Daily per capita protein supply: Animal Product (grams)	x_{20}	Accessibility and Utilization
	Daily per capita protein supply: Vegetal Product (grams)	x_{21}	Accessibility and Utilization
Social	Food supply (kilocalories per capita per day)	x_{22}	Utilization
	Food supply (protein grams per capita per day)	x_{23}	Utilization
	Food supply (fat grams per capita per day)	x_{24}	Utilization
	Percentage of urban population (%)	x_{25}	Stability
	Percentage of rural population (%)	x_{26}	Utilization
	Percentage of urbanization (%)	x_{27}	Stability
	Population	x_{28}	Accessibility

Step 10: Evaluate the forecasting accuracy of the trained OLS-MLR predictive mode by computing GoF measures (Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE)), comparing $\mathbf{Y}_{\text{obs}}$ and predicted rice production $\left(\mathbf{Y}_{\text{pred}} = \left[\hat{y}_i\right]_{n\times 1}\right)$, as expressed in Eqs. (7)-(9).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{7}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{8}$$

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{9}$$

Step 11: Repeat Steps 1-10 utilizing alternative feature selection performance metrics, including C_p and BIC, to assess the robustness of the selection process and identify the best-performing selection metrics.

Step 12: Rank the superiority of the feature selection method, including the deterministic feature selection method in the literature by utilizing the modified Taguchi-based VIKOR MCDM algorithm (Chuan *et al.* 2020).

Step 13: Deploy the superior OLS-MLR predictive model, incorporating the best set of determinants, for 5-year ahead short-run forecasting.

To address financial constraints, this study utilizes open-source secondary datasets. However, acquiring key determinants–such as government subsidies, farmer incomes, education levels, fertilizer prices and variants, rice prices, sea level rise, solar radiation, disease emergence, and weed control–was challenging due to data unavailability. Additionally, L-moments are employed instead of traditional statistical numerical summaries, as they offer greater robustness against outliers and are more suitable for limited datasets. Furthermore, statistical hypothesis tests are utilized for diagnostic checks of the OLS-MLR predictive model's assumptions, as they provide objective statistical evidence, whereas graphical visualizations frequently rely on subjective interpretation.

Meanwhile, a modified Taguchi-based VIKOR MCDM algorithm is employed to resolve conflicts between criteria, such as the number of determinants in the final set, and GoF measures. This study enhances the application of the MCDM algorithm by deviating from the approach in Chuan *et al.* (2020), which incorporated a Signal-to-Noise Ratio (SNR) transformation before applying the VIKOR MCDM algorithm.

4. Analysis, Results and Discussions

This section presents the analysis results utilizing R statistical software, while Microsoft Excel is utilized for storing datasets from multiple open-source databases, and for visualization and graphing. All analyses were conducted on a mid-end computing environment (Intel(R) Core (TM) i5-10210U CPU @ 1.60GHz, 4 Core(s), 8 Logical Processor(s)). To achieve the study's principal objective, the analysis follows the CRISP-DM framework. Specifically, Section 4.1 integrates the first three phases—Business Understanding, Data Understanding, and Data Preparation–focusing on descriptive and Exploratory Data Analysis (EDA). Meanwhile, Section 4.2 incorporates the final three phases–Modeling, Evaluation, and Deployment–emphasizing inferential analysis, predictive modeling, and 5-year ahead short-run forecasting deployment utilizing the best set of determinants.

4.1. *Analysis results of business understanding, data understanding, and data preparation*

The primary business objective of this article is to provide valuable insight for academicians, policymakers, smallholder farmers, and society by introducing a more effective feature selection method. This approach aims to enhance decision-making in policy development and

farming practices, ultimately strengthening food security and sustainable agriculture strategies. In line with this, the primary data mining objective is to propose an innovative hybrid deterministic best subset feature selection method to identify significant determinants impacting rice production in Malaysia. Based on these selected determinants, a parsimonious OLS-MLR predictive model is developed, ensuring a more efficient and interpretable approach to agricultural decision-making.

The implementation of feature selection requires a comprehensive foundation in data understanding, including examining the association between an endogenous variable and determinants, summarizing dataset characterizations, and conducting EDA. Based on the correlation coefficient classification (Chuan *et al.* 2023), analysis from the correlogram and correlation matrix consistently reveals that $\mathbf{X_4}$, $\mathbf{X_9}$, $\mathbf{X_{12}}$, and $\mathbf{X_{14}}$ exhibit low association with **Y**. Additionally, several determinants ($\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$, $\mathbf{X_5}$, $\mathbf{X_7}$, $\mathbf{X_8}$, $\mathbf{X_{10}}$, $\mathbf{X_{13}}$, $\mathbf{X_{16}}$, $\mathbf{X_{17}}$, $\mathbf{X_{18}}$, $\mathbf{X_{19}}$, $\mathbf{X_{20}}$, $\mathbf{X_{22}}$, $\mathbf{X_{23}}$, $\mathbf{X_{24}}$, $\mathbf{X_{25}}$, $\mathbf{X_{26}}$, $\mathbf{X_{27}}$, and $\mathbf{X_{28}}$) display high association with one another, indicating potential multicollinearity. However, the low association between **Y** and certain $\mathbf{X}_j$, as well as the high association among $\mathbf{X}_j$, does not necessarily imply that these $\mathbf{X}_j$ must be eliminated before implementing feature selection in the OLS-MLR predictive model. This is due to the potential adjustment effect, where an $\mathbf{X}_j$ that appears low association may contribute significantly when considered in the presence of other correlated variables. As a result, all the determinants presented in Table 1 are retained for applying the proposed innovative hybrid deterministic best subset feature selection method.

Additionally, Table 2 presents the EDA analysis results, including L-moments numerical summaries (L-Means, L-CV, L-Skewness, and L-Kurtosis), mild and extreme outlier detection, and the normality assessment of each variable utilizing the Shapiro-Wilk test. Statistically, L-Means function similarly to arithmetic means in characterizing the central tendency of each variable. However, L-moments are more robust to outliers, making them particularly useful for describing datasets with heavy tails. Based on Table 2, $\mathbf{X_{13}}$ and $\mathbf{X_{10}}$ exhibit the highest and lowest average values, respectively, across all variables. Despite that, direct comparisons of these average values are not meaningful, as these variables are measured in different units.

In contrast, L-CV characterizes the relative variation across variables, allowing values in Table 2 to be expressed as percentages. Among the 29 variables analyzed in this study, $\mathbf{X_{13}}$, and $\mathbf{X_2}$ exhibit the highest and lowest relative variation, respectively. The high L-CV of $\mathbf{X_{13}}$ suggests substantial year-over-year variability, which may be influenced by factors such as geographical conditions, climate change, national and international economic strategies and policies, and agriculture regulations. Conversely, Malaysia's tropical rainforest Köppen climate is characterized by relatively consistent high temperatures and rainfall throughout the year. This climatic stability may contribute to the low relative variation observed for $\mathbf{X_2}$.

L-Skewness and L-Kurtosis characterize the asymmetry and peakedness of a distribution. Statistically, these two numerical summaries aid in describing the shape of the distribution for each variable, where values approaching zero frequently suggest normality based on conventional guidelines in the literature. However, such guidelines frequently involve subjective decision-making and may not always be appropriate. In this study, although L-Skewness and L-Kurtosis for all variables in Table 2 are remarkably close to zero, the Shapiro-Wilks test indicates that most datasets are not normally distributed, except for **Y**, $\mathbf{X_2}$, $\mathbf{X_4}$, and $\mathbf{X_{21}}$. These findings highlight not merely the limitation of relying on subjective normality guidelines but also reinforce the advantages of L-moments over traditional numerical summaries in distributional analysis.

Table 2: Exploratory data insights: endogenous variable and determinants

Variable	L-Moments				Outlier(s)		Normality
	L-Means	L-CV	L-Skewness	L-Kurtosis	Mild	Extreme	
Y	2000855.1475	0.1260	-0.0706	0.1213	No	No	Yes
$\mathbf{X}_1$	4.1774	0.3598	0.0673	-0.0834	No	No	No
$\mathbf{X}_2$	30.2666	**0.0068**	0.0387	0.0936	No	No	Yes
$\mathbf{X}_3$	21.9289	0.0117	-0.0409	0.0080	No	No	No
$\mathbf{X}_4$	2931.0103	0.0552	0.0458	0.0554	No	No	Yes
$\mathbf{X}_5$	44.8716	0.3392	0.1652	0.0819	No	No	No
$\mathbf{X}_6$	22.5028	0.1794	0.0330	0.0565	No	No	No
$\mathbf{X}_7$	70.3979	0.3633	0.0005	-0.0196	No	No	No
$\mathbf{X}_8$	414.8188	0.2382	0.1161	-0.0642	No	No	No
$\mathbf{X}_9$	3.5755	0.4723	-0.2437	0.2927	Yes	Yes	No
$\mathbf{X}_{10}$	**0.0415**	0.2382	0.1161	-0.0642	No	No	No
$\mathbf{X}_{11}$	64.4569	0.2965	0.0827	-0.0240	No	No	No
$\mathbf{X}_{12}$	6.1155	0.2933	-0.2142	0.2125	Yes	Yes	No
$\mathbf{X}_{13}$	**73859500000.0000**	**0.5482**	0.2550	-0.0500	No	No	No
$\mathbf{X}_{14}$	2.9546	0.4740	0.2352	0.2628	Yes	Yes	No
$\mathbf{X}_{15}$	133.9292	0.1880	0.1411	-0.0041	No	No	No
$\mathbf{X}_{16}$	202.8259	0.5034	0.3062	0.1266	Yes	No	No
$\mathbf{X}_{17}$	3070377.0492	0.2699	0.2439	0.0133	No	No	No
$\mathbf{X}_{18}$	2804.7457	0.0251	-0.1255	0.1275	Yes	No	No
$\mathbf{X}_{19}$	81.3488	0.1033	-0.1859	0.0905	No	No	No
$\mathbf{X}_{20}$	35.4544	0.2149	-0.0282	-0.0061	No	No	No
$\mathbf{X}_{21}$	34.8409	0.0398	-0.0835	0.1767	Yes	No	Yes
$\mathbf{X}_{22}$	2804.7457	0.0251	-0.1255	0.1275	Yes	No	No
$\mathbf{X}_{23}$	70.2947	0.1141	0.0550	-0.0120	No	No	No
$\mathbf{X}_{24}$	81.3488	0.1033	-0.1859	0.0905	No	No	No
$\mathbf{X}_{25}$	52.4965	0.1776	0.0136	-0.0383	No	No	No
$\mathbf{X}_{26}$	47.5035	0.1963	-0.0136	-0.0383	No	No	No
$\mathbf{X}_{27}$	52.4965	0.1776	0.0136	-0.0383	No	No	No
$\mathbf{X}_{28}$	19291414.3770	0.2427	0.0947	-0.0360	No	No	No

*Note: L-CV signified the L-Coefficient of Variation

On the other hand, the presence of extreme outliers can impact the accuracy and reliability of statistical inference, including parameter estimation in the OLS-MLR predictive model. To address this issue, this study employed the 1.5IQR and 3IQR rules to detect mild and extreme outliers. As shown in Table 2, 7 out of 29 variables ($\mathbf{X_9}$, $\mathbf{X_{12}}$, $\mathbf{X_{14}}$, $\mathbf{X_{16}}$, $\mathbf{X_{18}}$, $\mathbf{X_{21}}$, and $\mathbf{X_{22}}$) contain mild outliers, while merely 3 variables ($\mathbf{X_9}$, $\mathbf{X_{12}}$, and $\mathbf{X_{14}}$) contain extreme outliers. Despite this, extreme outliers in time-series datasets are not removed but can be corrected utilizing capping methods. However, in this study, extreme outlier correction was not performed before feature selection. Moreover, from a statistical modeling perspective, an

extreme outlier does not necessarily imply a high-influence observation that can significantly impact inferential analysis results. Instead, these findings further validate the suitability of L-moments over traditional statistical summaries in characterizing datasets.

Understanding dataset characterization is essential for managing information redundancy, optimizing computational resources, improving predictive performance, and simplifying interpretation. Table 3 presents the feature selection analysis results, including performance metrics and the number of selected determinants across six hybrid (A7-A12) and six non-hybrid feature selection methods. Specifically, A1, A2, and A3 represent forward selection, backward elimination, and stepwise wrapper feature selection methods, respectively. Meanwhile, A4, A5, and A6 correspond to the best subset wrapper feature selection methods based on Adj-R^2, C_p, and BIC performance metrics, respectively.

Table 3: Feature selection analysis: performance metrics and number of selected determinants

Method	Mechanism	Number of Determinants	Performance Metrics			
			AIC	Adj-R^2	Cp	BIC
$A1^F$	Incremental	23	1385.6700	0.9717	-	-
$A2^F$	Decremental	14	1372.1600	0.9756	-	-
$A3^F$	Incremental & Decremental	14	1372.1600	0.9756	-	-
$A4^F$	Incremental	23	-	0.9750	-	-
$A5^F$	Incremental	22	-	-	19.8953	-
$A6^F$	Incremental	22	-	-	-	-155.7559
$A7^{F,*}$	Hybrid	11	-	0.9741	-	-
$A8^{F,*}$	Hybrid	11	-	0.9741	-	-
$A9^{F,*}$	Hybrid	11	-	0.9741	-	-
$A10^F$	Hybrid	11	-	0.9741	-	-
$A11^F$	Hybrid	11	-	0.9741	-	-
$A12^F$	Hybrid	11	-	0.9741	-	-

Note: "F" signified the approach is statistically significant based on the overall significance *F*-test; "*" signified the approach proposed in the literature, where A1—forward wrapper feature selection; A2—backward elimination wrapper feature selection; A3—stepwise wrapper feature selection; A4—best subset wrapper feature selection with Adj-R^2; A5—best subset wrapper feature selection with Cp; A6—best subset wrapper feature selection with BIC; A7—hybrid forward wrapper and Student's *t*-test filter feature selection; A8—hybrid backward elimination wrapper and Student's *t*-test filter feature selection; A9—hybrid stepwise wrapper and Student's *t*-test filter feature selection; A10—hybrid best subset wrapper and Student's *t*-test filter feature selection (Adj-R^2); A11—hybrid best subset wrapper and Student's *t*-test filter feature selection (Cp); A12—hybrid best subset wrapper and Student's *t*-test filter feature selection (BIC).

In contrast, A7, A8, and A9 represent hybrid forward selection, backward elimination, and stepwise wrapper feature selection methods, integrating Student's *t*-test as a filtering feature selection method. Similarly, A10, A11, and A12 denote the proposed innovative hybrid best subset wrapper feature selection methods, combining Adj-R^2, Cp, and BIC performance metrics with Student's *t*-test filtering. Among these methods, A7, A8, and A9 have been previously applied in the literature, although discussion on deterministic feature selection methods in agricultural economics and engineering research remains limited. In contrast, A1–A6 were included in this study for comparative analysis to further assess the effectiveness of the proposed innovative hybrid wrapper-filter feature selection methods.

As described in Section 3.2, the *F*-test evaluates the overall significance of the determinant set but does not assess the significance of individual determinants, which does not align with the principle of parsimony. This observation is further supported by the analysis results depicted in Table 3. Specifically, A1–A6 non-hybrid deterministic feature selection methods frequently resulted in higher selected determinants than the A7–A12 hybrid deterministic

feature selection methods. To further validate the effectiveness of the proposed hybrid deterministic feature selection methods, these selected determinant sets are incorporated into the OLS-MLR predictive model for forecasting accuracy evaluation, as detailed in the subsequent section.

4.2. *Analysis results of modeling, evaluation, and deployment*

The performance metrics in Table 3 are not suitable for determining the superiority of the hybrid and non-hybrid deterministic feature selection methods. Therefore, this study employed the well-known OLS-MLR predictive model, widely utilized in rice production prediction, to evaluate the feature selection methods. The analysis results, including GoF measures such as RMSE, MAE, and MAPE, are shown in Table 4. Forecasting accuracy was evaluated utilizing an internal validation approach instead of the hold-out cross-validation method, which was excluded due to the limited sample size. The hold-out cross-validation method could cause overfitting and unreliable results in a small dataset. Instead, the internal validation treats the test set as a proxy for future datasets, though this approach may not fully capture the challenges of forecasting future conditions. This study also employs the modified Taguchi-based VIKOR MCDM algorithm to rank methods, considering both the number of determinants and GoF measures. Despite adjusting the ranking direction to avoid irrational analysis results, inconsistencies remain, especially when similar feature sets (A7–A12) produce different rankings, pointing to the limitations of the MCDM approach.

Table 4: Forecasting accuracy evaluation utilizing A1–A12 feature selection methods

Method	**Forecasting accuracy evaluation**			**Rank**
	RMSE	MAE	MAPE	
A1	59581.1584	45281.8397	2.4228	7
A2	59951.1363	45230.7970	2.4508	7
A3	59951.1363	45230.7970	2.4508	7
A4	57893.5065	44157.4350	2.3745	7
A5	58504.0874	45095.1430	2.4133	7
A6	58504.0874	45095.1430	2.4133	7
A7*	63718.8384	49570.2855	2.6134	3
A8*	63718.8384	49570.2855	2.6134	3
A9*	**63718.8384**	**49570.2855**	**2.6134**	**2**
A10	63718.8384	49570.2855	2.6134	5
A11	63718.8384	49570.2855	2.6134	5
A12	**63718.8384**	**49570.2855**	**2.6134**	**1**

Note: "*" signified the approach proposed in the literature.

To strengthen the analysis results, a diagnostic check was performed on the OLS-MLR predictive model assumptions utilizing the best determinant set in Eq. (10). The analysis results showed violations of the independence (p-value for Run test: 0.0390), homoscedasticity (p-value for Breusch-Pagan test: 0.0061), and multicollinearity assumptions, although the normality assumption (p-value for Shapiro-Wilk test: 0.4580) was not violated. Despite these issues, the prediction accuracy remained largely unaffected, as shown in Figure 1. While these violations could impact inferential analysis, this study does not delve into detailed interpretations of the estimated parameters. The findings suggest that

the four key SDG pillars–environmental, economic, human, and social–have both favorable and detrimental impacts on annual rice production in Malaysia. Additionally, the significant intercept in Eq. (10) indicates the presence of unconsidered determinants, reinforcing the challenges posed by the dataset, as discussed in Section 3.3.

$$\begin{aligned}\hat{y} = {} & -4.448\text{E}6 + 8.199\text{E}4x_1 + 4.528\text{E}3x_8 - 1.990\text{E}{-}6x_{13} + 9.995\text{E}3x_{14} \\ & - 1.360\text{E}3x_{15} + 7.712\text{E}2x_{16} - 3.537\text{E}{-}1x_{17} + 2.830\text{E}6x_{20} + 2.820\text{E}6x_{21} \\ & - 2.797\text{E}6x_{23} + 6.655\text{E}4x_{25}\end{aligned} \tag{10}$$

In summary, the analysis results presented in Table 4 show that the hybrid deterministic feature selection methods consistently outperformed the non-hybrid deterministic feature selection methods. The hybrid feature selection methods yielded parsimonious, computationally efficient, and interpretable predictive models. These findings also highlight the effectiveness of the proposed innovative hybrid deterministic feature selection methods, which performed comparably to the hybrid feature selection methods discussed in the literature, such as those by Chuan *et al.* (2024). However, this study advances their work by incorporating potential determinants across all four key pillars of the SGDs and considering the dimension of food security.

This study emphasizes that although A7–A12 methods yielded identical forecasting accuracy metrics and converged on a similar final set of significant determinants following hybrid filtering, the ranking variations in Table 4 stem from the normalization and weighted aggregation processes intrinsic to the modified Taguchi-based VIKOR MCDM algorithm. Minor numerical differences during intermediate steps can lead to subtle variations in the overall utility and regret measures, which in turn influence the final ranking. These differences reflect the methodological sensitivity of MCDM-based rankings, particularly due to the transformation of forecasting accuracy metrics into SNR before normalization and aggregation, and should not be interpreted as an error in the performance evaluation.

To further validate the effectiveness of the proposed hybrid deterministic wrapper-filter feature selection method, this study applied the best-selected determinant set within the OLS-MLR predictive model to predict annual rice production from 1961 to 2021, followed by a 5-year ahead short-run forecasting, as illustrated in Figure 1. The scatter diagram in Figure 1 shows that $\hat{\mathbf{Y}}_i$ (gold-colored dots) closely align with $\mathbf{Y}_i$ (dark blue-colored dots), with both exhibiting nearly identical linear trend lines. This further supports the argument that the violations of OLS-MLR predictive model assumptions did not significantly impact predictive performance. Moreover, the 5-year ahead short-run forecasting results (dark red-colored dots) are consistent with the mild evidence reported by the Department of Statistics Malaysia (2022). However, a deviation of y_i and $\hat{y}_i$ in 2022 was observed (Department of Statistics Malaysia 2023), primarily attributed to uncontrollable factors such as an insufficient supply of basic paddy seeds and prevailing farming practices (Kasinathan 2023). These factors are massively impacted by broader agricultural policies and external conditions beyond the scope of this study. Despite uncertainties regarding formal statistics on paddy and y_i beyond 2022, the observed trend suggests a potential continuation of growth. This forecasted outcome aligns with Malaysia's target of achieving a Self-Sufficiency Ratio (SSR) of 73.8% in 2022 and 80% by 2030 (Kasinathan 2023), reinforcing the relevance of this study's findings in the context of national agriculture goals.

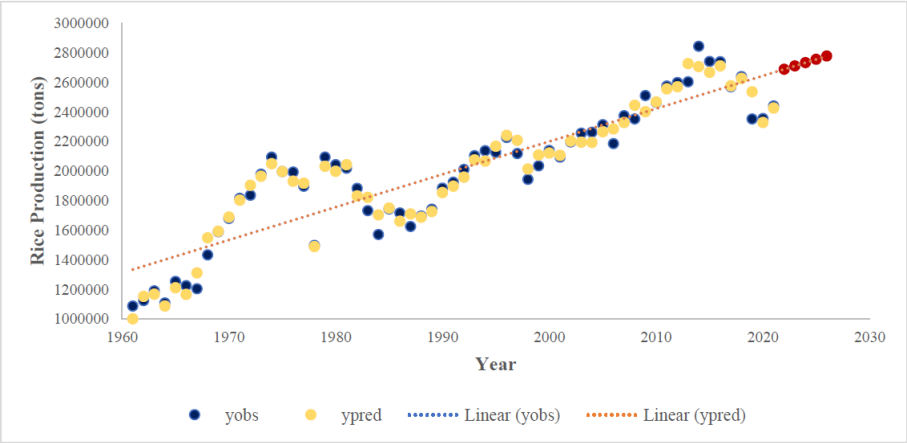

Figure 1: Comparison of the $\mathbf{Y}_i$ (yobs) and $\hat{\mathbf{Y}}_i$ (ypred) values based on the best set of determinants utilizing the OLS-MLR predictive model

5. Conclusion and Future Work

This study proposed an innovative hybrid deterministic wrapper-filter feature selection method for identifying key determinants impacting rice production in Malaysia utilizing the OLS-MLR predictive model. The proposed hybrid wrapper-filter approach integrates the best subset method with Student's *t*-test filtering, evaluated utilizing Adj-R^2, C_p, and BIC performance metrics. To address the conflicts among selection criteria, such as the number of determinants and GoF measures, a modified Taguchi-based VIKOR MCDM algorithm was employed for ranking feature selection methods. However, findings revealed limitations in MCDM rankings, particularly when criteria with similar values received different rankings. Despite this, analysis results confirmed that hybrid deterministic feature selection methods consistently outperformed non-hybrid deterministic feature selection methods, yielding parsimonious, computationally efficient, and interpretable predictive models.

Furthermore, this study found that violations of OLS-MLR assumptions (independence, homoscedasticity, and multicollinearity) did not impact prediction accuracy, though they impacted inferential analysis. Consequently, the estimated regression parameters interpretation was not detailed. To address this limitation, future studies should explore the Bayesian parameter estimation method and multivariate predictive models to enhance interpretability, particularly in the context of SDGs and food security. Overall, this research provides valuable insights for academicians, policymakers, smallholder farmers, and society, offering a more effective feature selection method that supports policy development and farming practices. Additionally, it contributes to both academia and industry realms in introducing a hybrid deterministic features selection method with improved practical application and interpretability compared to stochastic metaheuristic approaches. By bridging methodological advancements with practical applications, this study contributes to more informed decision-making in food security and sustainable agriculture.

Acknowledgments

The author sincerely appreciates the Climate Change Knowledge Portal (CCKP), Our World in Data (OWID), and the World Bank for providing open-source secondary datasets, which were instrumental in completing this study. The author also extends gratitude to Universiti

Malaysia Pahang Al-Sultan Abdullah for funding this research through the Fundamental Research Grant UMPSA (Grant No.: RDU220393).

References

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., Talib B. & Toriman M.E. 2011. The relationships between the socio-economic profile of farmers and paddy productivity in North-West Selangor, Malaysia. *Asia-Pacific Development Journal* **18**(1): 161-173.

Alam M.M., Siwar C., Talib B. & Toriman M.E. 2014. Impacts of climatic changes on paddy production in Malaysia: Micro study on IADA at North West Selangor. *Research Journal of Environmental and Earth Sciences* **6**(5): 251-258.

Cheng A. 2023. Evaluating Fintech industry's risks: A preliminary analysis based on CRISP-DM framework. *Finance Research Letters* **55**(Part B): 103966.

Chuan Z.L., Deni S.M, Fam S.F. & Ismail N. 2020. The effectiveness of a probabilistic principal component analysis model and expectation maximisation algorithm in treating missing daily rainfall data. *Asia-Pacific Journal of Atmospheric Sciences* **56**: 119-129.

Chuan Z.L., Fam S.F., Lee Q.H., Kok J.S. & Azam M.N.B.M. 2022. Modeling 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* **3**(2): 1-12.

Chuan Z.L., Sheng T.R., Cheng T.C., Sern A.L.B., Luen D.L.K. & Sai C.Y. 2025. Smart agriculture economics and engineering: Unveiling the innovation behind AI-enhanced rice farming. *Multidisciplinary Applied Research and Innovation*. *Multidisciplinary Applied Research and Innovation* **6**(2): 1-17.

Chuan Z.L., Japashov N., Yuan S.K., Qing T.W. & Ismail N. 2024a. Analyzing enrolment patterns: Modified stacked ensemble statistical learning-based approach to educational decision-making. *Akademika* **94**(2): 232-251.

Chuan Z.L., Tan L.K., Chyin A.W.C., Tham Y.H., Ong S.J., Low J.Y. & Sai C.Y. 2024b. Sustainable energy management: Artificial intelligence-based electricity consumption prediction in limited dataset environment for industry applications. *Matematika* **40**(3): 143-167.

Chuan Z.L., Wei D.C.T., Aminuddin A.S.B.A., Fam S.-F. & Ken T.L. 2024c. Comparison of multiple linear regression and multiple nonlinear regression models for predicting rice production. *AIP Conference Proceedings* **3150**(1): 050008.

Chuan Z.L., Wei D.C.T., Yan C.L.W., Nasser M.F.A., Ghani N.A.M., Jemain A.A. & Liong C.Y. 2023. A comparative of two-dimensional statistical moment invariants features in formulating an automated probabilistic machine learning identification algorithm for forensic application. *Malaysian Journal of Fundamental and Applied Sciences* **19**(4): 525-538.

Department of Statistics Malaysia. 2022. *Selected agricultural indicators, Malaysia, 2021*. https://www.dosm.gov.my/portal-main/release-content/selected-agricultural-indicators-malaysia-2021 (20 January 2025).

Department of Statistics Malaysia. 2023. *Selected agricultural indicators, Malaysia, 2023*. https://www.dosm.gov.my/portal-main/release-content/selected-agricultural-indicators-malaysia- (20 January 2025).

Economic Planning Unit Prime Minister's Department. 2021. *Twelfth Malaysia Plan 2021-2025: A prosperous, inclusive, sustainable Malaysia*. Kuala Lumpur: Percetakan Nasional Malaysia Berhad.

Fauzi F.D. & Bakar A.S.A. 2022. Rice production forecasting in Malaysia: A Box-Jenkins and ARIMA model approach. *Proceedings of the 8th Annual ECOFI Symposium 2022*, pp. 212-220.

Food and Agriculture Organization of the United Nations. 2024. FAOSTAT: Suite of food security indicators. https://www.fao.org/faostat/en/#data/FS (25 July 2024).

Garcia-Arteaga J., Zambrano-Zambrano J., Parraga-Alava J. & Rodas-Silva J. 2024. An effective approach for identifying keywords as high-quality filters to get emergency-implicated Twitter Spanish data. *Computer Speech & Language* **84**:101579.

Gunaratne M.S., Firdaus R.B.R. & Rathnasooriya S.I. 2021. Climate change and food security in Sri Lanka: Towards food sovereignty. *Humanities & Social Sciences Communications* **8**: 229.

Hosking J.R.M. 1990. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* **52**(1): 105-124.

Idalisa N., Rivaie M., Fadhilah N.H., Atikah N., Shahida A. & Noh N.H.M. 2019. Multiple linear regression model of rice production using conjugate gradient methods. *Matematika* **35**(2): 229-235.

Kasinathan S. 2023. National audit report shows nearly quarter of Malaysian paddy farmers earn below RM600 monthly, as rice cultivation programme fails to reach target. *Malay Mail*.

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 (20 January 2025).

Liang C.Z., Sern A.L.B., Cheng T.C., Luen D.L.K., Japashov N. & Hiae T.E. 2024. Empowering Industry 5.0: Nurturing STEM tertiary education and careers through Additional Mathematics. In Al-Humairi S.N.S. (ed.). *Utilizing Renewable Energy, Technology, and Education for Industry 5.0*: 124-155. IGI Global.

Lohaj O., Paralič J., Bednár P., Paraličová Z. & Huba M. 2023. Unraveling COVID-19 dynamics via machine learning and XAI: Investigating variant influence and prognostic classification. *Machine Learning & Knowledge Extraction* **5**(4): 1266-1281.

Makhtar S., Zaidin Abidin I.S. & Islam R. 2022. The impact of rice productivity in Malaysia: Econometric analysis. *International Journal of Business and Economy* **4**(3): 21-32.

Manjunath M.C. & Pallayan B.P. 2024. Artificial Bee Colony algorithm-based feature selection and hybrid ML framework for efficient rice yield prediction. *International Journal of Electrical and Computer Engineering Systems* **15**(3): 235-246.

Marong M., Husin N.A., Zolkepli M. & Affendey L.S. 2024. Data-driven rice yield predictions and prescriptive analytics for sustainable agriculture in Malaysia. *International Journal of Advanced Computer Science and Applications* **15**(3): 362-370.

Maya Gopal P.S. & Bhargavi R. 2019. Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms. *Applied Artificial Intelligence* **33**(7): 621-642.

Ministry of Economy. 2023. *Executive Summary Mid-Term Review Twelfth Malaysia Plan 2021-2025 (Malaysia Madani: Sustainable, Prosperous, High Income)*. Kuala Lumpur: Percetakan Nasional Malaysia Berhad.

Mishra S., Mishra D., Mallick P.K., Santra G.H. & Kumar S. 2021. A novel Borda count based feature ranking and feature fusion strategy to attain effective climatic features for rice yield prediction. *Informatica* **45**(1): 13-31.

Putri R.E., Yahya A., Adam N.M. & Abd Aziz S. 2019. Rice yield prediction model with respect to crop healthiness and soil fertility. *Food Research* **3**(2): 171-176.

Roslan N.M., Shinyie W.L. & Ling S.S. 2021. Modelling high dimensional paddy production data using Copulas. *Pertanika Journal of Science and Technology* **29**(1): 263-284.

Sathya P. & Gnanasekaran P. 2023. Ensemble feature selection framework for paddy yield prediction in Cauvery Basin using machine learning classifiers. *Cogent Engineering* **10**(2): 2250061.

Shahidan M**.S.**, Fatah F.A. & Lim H.E. 2022**.** Econometric modelling for estimating of paddy yield and rice production in Melaka, Malaysia. *IOP Conference Series: Earth and Environmental Science* **1059**: 012078.

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.

Wijayanti E.B., Setiadi D.R.I.M. & Setyoko B.H. 2024. Dataset analysis and feature characteristics to predict rice production based on eXtreme gradient boosting. *Journal of Computing Theories and Applications* **1**(3): 299-310.

Wong L., Kam A., Esa Z.M. & Kassim Q. 2024. Malaysia's long-term food security: The path beyond self-sufficiency ratios and import-dependent ratios. Policy Brief. Institute of Strategic & International Studies (ISIS) Malaysia.

Yusof Z.M., Misiran M., Baharin N.F., Yaacob M.F., Aziz N.A.B.A. & Sanan N.H.B. 2019. Projection of paddy projection in Kedah Malaysia: A case study. *Asian Journal of Advances in Agricultural Research* **10**(3): 1-6.

Statistics & Data Analytics Research Cluster
Center for Mathematical Sciences
Universiti Malaysia Pahang Al-Sultan Abdullah
Lebuh Persiaran Tun Khalil Yaakob
26300 Kuantan
Pahang, MALAYSIA
E-mail: chuanzl@umpsa.edu.my[*]

Center for Mathematical Sciences
Universiti Malaysia Pahang Al-Sultan Abdullah
Lebuh Persiaran Tun Khalil Yaakob
26300 Kuantan
Pahang, MALAYSIA
E-mail: limbingsern123@gmail.com, luenxd0107@gmail.com, zecheng.t@gmail.com

Faculty of Industrial Management
Universiti Malaysia Pahang Al-Sultan Abdullah
Lebuh Persiaran Tun Khalil Yaakob
26300 Kuantan
Pahang, MALAYSIA
E-mail: renshengtham@gmail.com

Received: 22 January 2025
Accepted: 25 May 2025

*Corresponding author