Data-Driven Rice Yield Predictions and Prescriptive Analytics for Sustainable Agriculture in Malaysia

Muhammad Marong, Nor Azura Husin, Maslina Zolkepli, Lilly Suriani Affendey
Faculty of Computer Science and Information Technology,
Department of Computer Science, University Putra Malaysia, Selangor, Malaysia

***Abstract*—Maximizing rice yield is critical for ensuring food security and sustainable agriculture in Malaysia. This research investigates the impact of environmental conditions and management methods on crop yields, focusing on accurate predictions to inform decision-making by farmers. Utilizing machine learning algorithms as decision-support tools, the study analyses commonly used models—Linear Regression, Support Vector Machines, Random Forest, and Artificial Neural Networks—alongside key environmental factors such as temperature, rainfall, and historical yield data. A comprehensive dataset for rice yield prediction in Malaysia was constructed, encompassing yield data from 2014 to 2018. To elucidate the influence of climatic factors, long-term rainfall records spanning 1981 to 2018 were incorporated into the analysis. This extensive dataset facilitates the exploration of recent agricultural trends in Malaysia and their relationship to rice yield. The study specifically evaluates the performance of Random Forest, Support Vector Machine (SVM), and Neural Network (NN) models using metrics like Correlation Coefficient, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). Results reveal Random Forest as the standout performer with a Correlation Coefficient of 0.954, indicating a robust positive linear relationship between predictions and actual yield data. SVM and NN also exhibit respectable Correlation Coefficients of 0.767 and 0.791, respectively, making them effective tools for rice yield prediction in Malaysia. By integrating diverse environmental and management factors, the proposed methodology enhances prediction accuracy, enabling farmers to optimize practices for better economic outcomes. This approach holds significant potential for contributing to sustainable agriculture, improved food security, and enhanced economic efficiency in Malaysia's rice farming sector. Leveraging machine learning, the research aims to transform rice yield prediction into a proactive decision-making tool, fostering a resilient and productive agrarian ecosystem in Malaysia.**

Keywords—Rice yield prediction; sustainable agriculture; linear regression; support vector machine; artificial neural network; predictive analytics

I. Introduction

Agriculture plays a pivotal role in Malaysia, contributing significantly to the nation's economic development. It remains one of the primary sources of livelihood for a substantial portion of the population, and its modernization is essential to meet the growing demands of the country's expanding population. In Malaysia, a significant portion of the land is utilized for agricultural activities, addressing the food requirements of millions of people. As the agricultural landscape evolves, farmers are increasingly exploring opportunities to enhance productivity and achieve optimal returns on their investments. Traditionally, crop yield predictions were heavily reliant on a farmer's experience and understanding of specific land and crops [1] [2] [3]. However, with changing conditions and the pursuit of diversified crops, there arises a need for more comprehensive and accurate data to guide farmers in their decision-making process. Farmers are seeking more information about new crops and their potential profitability to make informed choices regarding crop selection and overall agricultural practices [3] [4]. Accurate estimation of crop performance under various environmental conditions can significantly improve farm productivity and financial outcomes. The global rice market is narrow, making it highly susceptible to market fluctuations caused by supply interruptions due to weather variations [5] [6] [7]. To safeguard its population from hunger, Malaysia has implemented a protectionist policy for its paddy and rice business [8] [9]. Policymakers in Malaysia require accurate crop yield forecasts to evaluate the benefits and drawbacks of imports and exports, strengthening the country's food supply and ensuring its security [10] [11]. Predicting agricultural productivity is a complex task due to the numerous interconnected factors involved. To address these challenges, the application of Data Science techniques has become increasingly crucial. In the modern era of agriculture, ensuring food security stands as a paramount challenge, especially within the context of Malaysia. The burgeoning population and the evolving landscape of climate dynamics render the traditional methods of crop yield prediction inadequate. The delicate balance between supply and demand, often influenced by climatic vagaries, can disrupt the stable rhythm of food production. Malaysia, with its dependency on rice as a staple, stands at a critical juncture in safeguarding its food security.

At present, major suppliers of Rice in Malaysia and its exporters are largely planted in Malaysian states, namely Johor, Kedah, Kelantan, Melaka, Negeri Sembilan, Pahang, Perak, Perlis, Pulau Pinang, Selangor, Terengganu, Sabah, and Sarawak, as shown in Fig. 1. These states are the major contributors to Malaysia's rice production, accounting for a significant share of over five hundred thousand metric tons annually, thus making up the majority of rice production in Malaysia. The entire process from planting the crop to harvest is closely intertwined with data collection and its intricate relationship with crop yield. These collected data play a pivotal role in training Machine Learning (ML) algorithms. These ML algorithms provide invaluable insights for farmers to estimate the profitability of their crops. Armed with this knowledge, farmers can make well-informed decisions regarding their crop investments and consider necessary modifications before the harvest, often leading to additional costs to secure a more bountiful yield.

Historically, rice yield prediction was anchored in the expertise of farmers, an art passed down through generations. Yet, the intricacies of today's climate patterns, characterized by unprecedented shifts and erratic behavior, demand a more sophisticated and data-driven approach. The heightened susceptibility of the rice market to supply disruptions caused by weather anomalies underscores the urgency to fortify the nation's agricultural resilience.

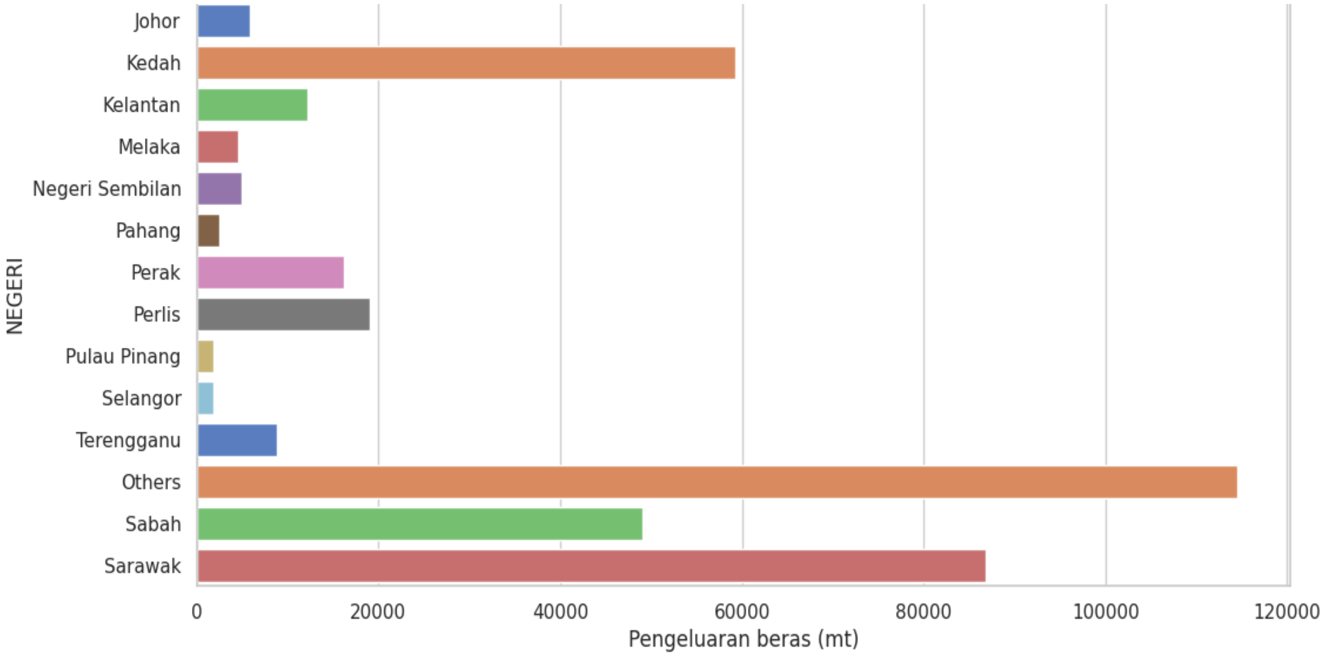


Fig. 1. Statistics of rice yield in Malaysia.

In this tapestry of challenges, the study illuminates a pivotal issue - the necessity for an advanced predictive model that not only harnesses the power of machine learning but also delves deeper into the complex interplay between environmental factors and crop productivity [12]. By addressing this issue, the research takes a bold step towards fortifying the foundations of food security in Malaysia.

Data analysis using scientific approaches can yield valuable insights into the data, enabling informed decision-making in agriculture. This study aims to explore the impact of environmental conditions and management practices on crop yields in Malaysia. By leveraging machine learning algorithms, the study seeks to develop accurate prediction models that can assist farmers in optimizing their agricultural practices and achieving better economic efficiency. This research holds the potential to contribute valuable insights into sustainable agriculture, improve food security, and enhance economic outcomes in Malaysia's agricultural sector.

The remaining part of the paper is organized as follows: In Section II, a comprehensive review of related works in the field of rice yield prediction in Malaysia using hybrid machine learning models is presented, emphasizing the significance of integrating diverse environmental and management factors for enhanced accuracy. Section III describes the materials and methods employed in this study, including the dataset used for rice yield prediction, and the implementation of hybrid machine learning models incorporating Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN). In Section IV, the study present the implementation details, results, and performance evaluation of the hybrid machine learning models. The evaluation includes metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared to assess the models' predictive capabilities. Additionally, a detailed discussion on the obtained results is provided, highlighting the strengths and limitations of the models. Lastly, in Section V, the conclusion of the study is presented, summarizing the key findings and potential future avenues for further research in the domain of rice yield prediction in Malaysia using hybrid machine learning models.

II. Literature Review

In Malaysia, estimating rice crop yields accurately is of utmost importance to ensure food security and support the nation's agricultural sustainability goals [23]. Traditional methods for rice yield prediction, such as static regression and mechanistic approaches, have limitations in effectively capturing the complex and nonlinear relationships between input factors, such as weather conditions, climate variations, and agricultural practices, and the resulting rice yields [23]. To address the challenges in rice yield prediction specifically for Malaysia, there is a pressing need to develop a hybrid machine learning and deep learning (ML/DL) model that can harness the power of data-driven techniques to enhance accuracy and predictive capabilities [13]. This hybrid approach should focus on leveraging historical yield data, weather information, and other pertinent features specific to Malaysia's rice-growing regions to create a robust and reliable predictive model [13][14].

The major aim of this research is to develop a hybrid machine learning and deep learning (ML/DL) model for accurate rice yield prediction in Malaysia. The model will integrate historical yield data, weather information, and relevant features specific to Malaysia's rice-growing regions. By autonomously extracting essential features, the model aims to achieve high prediction accuracy and generalization across different regions and cropping seasons. The practical implementation of the model will provide farmers with valuable insights for sustainable crop management practices contributing to Malaysia's food security and agricultural sustainability goals.

Estimating rice yields is critical in meeting the growing demand for food throughout the nation [38]. It aids in the enhancement of management procedures vital to maximizing agricultural yields. One of the challenges in rice yield prediction is the limited availability of diverse and comprehensive datasets [20]. To address this issue, researchers have been exploring the integration of remote sensing data, climate information, and historical crop yield data to train rice yield models effectively [15] [30]. In recent years, there has been a growing interest in developing hybrid models that combine both machine learning (ML) and deep learning (DL) techniques to enhance rice yield prediction accuracy [13][14]. The integration of ML algorithms with DL architectures employs the strengths of both approaches, resulting in more robust and accurate predictions [14]. Several studies have explored the benefits of hybrid ML/DL models for rice yield prediction.

One example of a successful hybrid ML/DL model is the combination of fuzzy clustering and DL for rice yield prediction [35]. Pham et al. (2021) developed a hybrid model that utilized weather data and satellite-based spectral indices for rice yield prediction in Vietnam. They first applied fuzzy clustering to categorize regions with similar environmental characteristics, creating distinct clusters. Next, they used deep learning techniques, such as adaptive neuro-fuzzy inference systems (ANFIS), within each cluster to predict rice yields. The hybrid model achieved a high accuracy of over 97%, outperforming individual ML or DL models [35]. Another successful hybrid model for rice yield prediction combined artificial neural networks (ANN) with support vector regression (SVR) [32]. Zhang et al. (2021) developed this hybrid model to predict rice yield in China. The ANN component captured the nonlinear relationships between input factors like weather, soil, and management practices, while SVR provided robustness and improved generalization. The hybrid model achieved a high prediction accuracy of over 90%, demonstrating the effectiveness of combining ML and DL techniques [32]. Hybrid ML/DL models can also integrate DL techniques with crop simulation models, offering insights into complex rice growth dynamics [33]. Sharma et al. (2021) developed a hybrid model that utilized remote sensing data and weather data to train a DL model. The output from the DL model was then used as input to a crop simulation model, which captured the interactions between environmental conditions and crop growth. This integrated approach achieved a high prediction accuracy of over 93% for rice yield prediction in India [33].

Moreover, hybrid models can also reduce the risk of overfitting and improve the generalizability of the models [33]. Overfitting occurs when the model is too complex and fits the training data too closely, resulting in poor performance on unseen data. Hybrid models can mitigate this risk by combining models with different biases and variance, resulting in a more balanced and robust model [33]. For example, they found that their hybrid model outperformed individual models in terms of generalization and robustness. The combination of deep learning techniques and crop simulation models allowed the model to capture the spatial and temporal variations in the rice growth conditions and yield output [33] [34].

In summary, hybrid machine learning models have shown promising results in rice yield prediction [13]. It gives ability to improve prediction accuracy, robustness, and generalization [14] [20]. The combination of machine learning techniques with other data-driven or statistical models can capture the complex relationships between the input factors and the yield output more accurately and reduce the risk of overfitting [15] [16]. These developments in the field have significant implications for enhancing food security and nutrition, particularly in developing countries where rice is a staple food crop [38][39].

III. Materials and Methods

A. *Case Study and Data Description*

Rice is a staple crop in Malaysia, and accurate yield prediction plays a crucial role in ensuring food security and optimizing agricultural practices. Traditional regression-based models and mechanistic approaches have limitations in capturing the complex and nonlinear relationships between crop yield and various influencing factors, including weather conditions and agricultural practices. To address these challenges, the study proposes a hybrid machine learning and deep learning (ML/DL) model for rice yield prediction in Malaysia. The model employs the advantages of both ML and DL techniques to enhance prediction accuracy and generalization. The data used in this study encompass historical rice yield records and weather data from multiple rice-growing regions in Malaysia. The historical yield data is collected over several cropping seasons, covering a significant time span to account for inter-annual variations. Daily weather information, including rainfall, temperature, and humidity, is obtained from meteorological stations situated in close proximity to the rice fields. Additionally, Flood data has been included in the dataset, depicted in Fig. 2.

B. *Challenges of Traditional Models and Hybrid Model Justification*

Traditional regression-based models and mechanistic approaches encounter several challenges when predicting rice yields. These challenges include limited flexibility, an inability to capture nonlinear patterns, assumptions of homoscedasticity, and overlooking intricate interactions between influencing factors. Traditional models often struggle to adapt to the dynamic and complex relationships inherent in rice production, leading to suboptimal predictions.

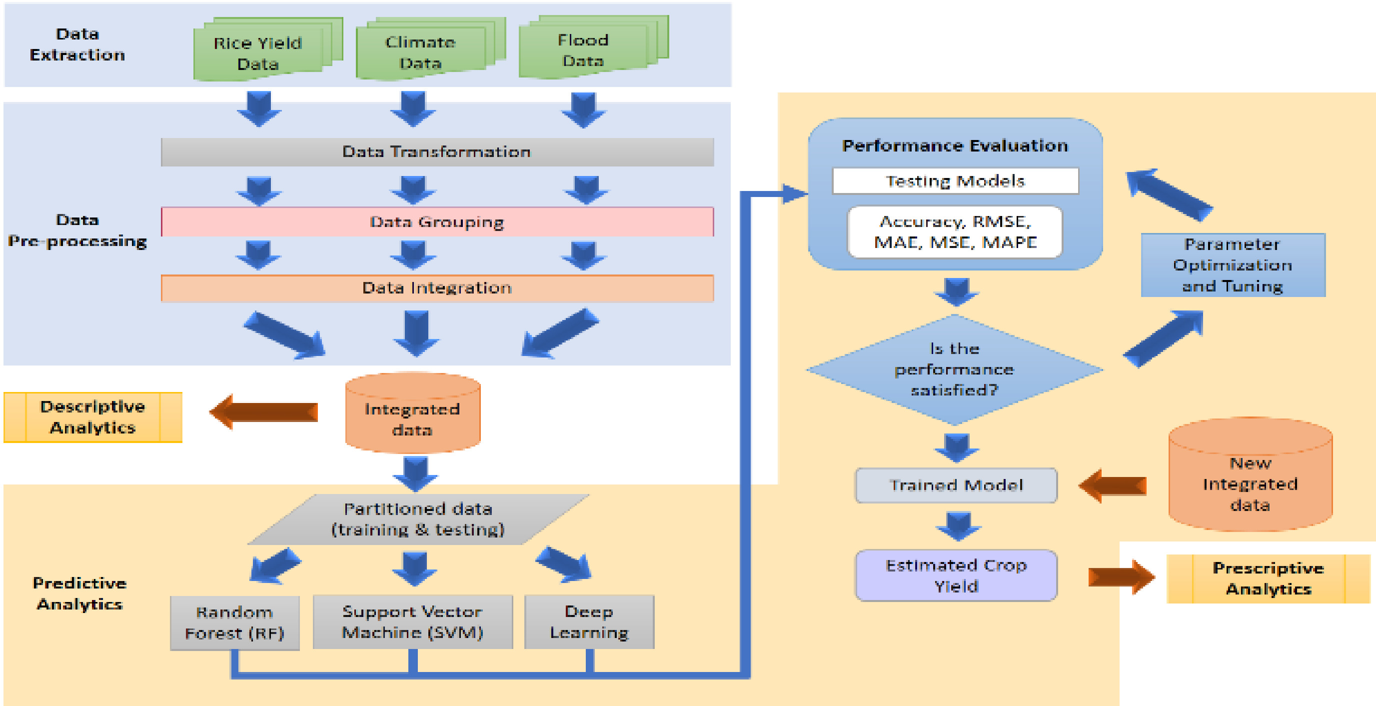


Fig. 2. Architecture of proposed model.

The proposed hybrid machine learning and deep learning (ML/DL) model offer a compelling solution to these challenges. By integrating RF, SVM, and ANN, the hybrid model leverages the strengths of each algorithm. RF efficiently handles non-linear relationships, capturing complex interactions within the dataset. SVM's non-linear kernel enhances adaptability to high-dimensional data, crucial for addressing intricate relationships in rice yield prediction. The deep learning capabilities of ANN allow the model to learn hierarchical representations, effectively capturing complex temporal patterns and dependencies within the dataset. In combination, these algorithms overcome the limitations of traditional models, providing a more robust framework for accurate and sustainable rice yield predictions in Malaysia. The hybrid model's flexibility, adaptability, and ability to capture nonlinear patterns make it well-suited for the challenges posed by the complex dynamics of rice production.

C. Models for Forecasting

In this study, the researcher employs three powerful algorithms for forecasting rice yield in Malaysia: RF, SVM, and ANN. Each model brings unique strengths to handle the complexity of climate data and its impact on rice production.

Random Forest, an ensemble learning method, constructs multiple decision trees during training and combines their outputs for accurate predictions [17]. Given the intricate relationships between climate variables and rice yield in Malaysia, RF is well-suited to efficiently handle non-linear patterns, capturing the complex interactions within the dataset [17][20]. The selection of RF aligns with the objective of sustainable agriculture by providing insights into feature importance, aiding the understanding of the impact of environmental conditions and management practices on rice yield.

SVM, a supervised learning algorithm, separates different yield classes based on climate and agricultural attributes. It finds the optimal hyperplane maximizing the margin between data points of different classes. SVM's ability to handle high-dimensional data and non-linear kernels makes it valuable for crop yield prediction [13] [21]. In the context of sustainable agriculture, SVM is chosen for its effectiveness in high-dimensional spaces and its capacity to handle non-linear relationships, essential for capturing the complexity of rice yield determinants. This aligns with the goal of optimizing agricultural practices.

Artificial Neural Network, a deep learning technique, is inspired by the human brain's neural networks. With interconnected layers of nodes, ANN processes input data and learns complex patterns [22]. Deep learning approaches, including ANN, have shown promising results in crop yield prediction [15] [19]. ANN is selected for its capability to learn hierarchical representations, making it suitable for capturing intricate relationships within the data. This is particularly valuable for sustainable agriculture, where understanding complex patterns and dependencies is crucial for informed decision-making.

For model evaluation, the hybrid ML/DL model combines the outputs of RF, SVM, and ANN. Cross-validation techniques assess model performance, and hyperparameter tuning optimizes their configurations. Metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared gauge predictive capabilities.

By integrating these algorithms, the hybrid ML/DL model effectively handles climate data complexities, providing accurate predictions of rice yield in Malaysia. Findings offer valuable insights to farmers and policymakers, aiding informed decisions for enhanced agricultural productivity and food security in the region.

D. Optimization Technique

The optimization of the hybrid ML/DL model is essential to ensure its effectiveness in forecasting rice yield in Malaysia accurately [36]. Hyperparameter tuning, a crucial optimization technique, is employed to fine-tune the parameters of the RF, SVM, and ANN algorithms [31]. Hyperparameters significantly impact the performance of these models, and their optimal selection is vital to achieve the best predictive results [29].

Grid search, a systematic hyperparameter tuning method, is utilized to explore various combinations of hyperparameter values within predefined ranges [30]. It exhaustively searches through the hyperparameter space and evaluates each combination's performance using cross-validation techniques [30]. The combination of hyperparameter values that yields the best performance metrics on the validation set is chosen as the optimal configuration for each model [30].

By fine-tuning the hyperparameters, the study aims to optimize the models for forecasting rice yield and improve their generalization capabilities [33]. The optimization process ensures that the models can effectively capture the complex relationships between climate variables and rice production, leading to accurate and reliable yield predictions for Malaysia's agricultural sector [33].

E. Architecture of the Model

In the proposed framework, machine learning and deep learning techniques are employed to predict the best crop production for a given dataset. The prediction of suitable crops is based on the analysis of current atmospheric and climatic parameters. These deep learning techniques excel in capturing complex temporal patterns and dependencies within the dataset, making them well-suited for crop prediction tasks. On the other hand, the RF and SVM algorithm is implemented under machine learning to handle classification tasks and assist in the crop prediction process. By combining the strengths of machine learning and deep learning, the proposed model can effectively provide precise and reliable crop predictions without the need for soil-related data.

F. Proposed Framework

The scenario below is intended for illustrative purposes only. The actual model predictions and results will be based on the real-world data and factors encountered during rice production in Malaysia.

Fig. 2 employs hybrid machine learning and deep learning techniques to predict rice yield in Malaysia. To demonstrate the model's performance under various conditions, the study presents a hypothetical scenario that highlights its ups and downs during a crop season. The study begins with an optimal climate phase characterized by moderate rainfall and abundant sunshine. During this phase, the hybrid ML/DL model accurately predicts a bountiful rice yield, effectively capturing the positive correlation between climatic factors and crop productivity [17] [35]. However, as the crop season progresses, unforeseen weather fluctuations, such as unexpected heavy rainfall and prolonged drought spells, challenge the model's adaptability to non-linear relationships between extreme events and their impact on yield [19][31]. As a result, there might be slight deviations in the model's predictions from the actual yield during this period [20] [22] [26]. Moreover, localized pest infestations during the mid-crop season further test the model's capabilities, as it primarily focuses on climate-related factors and faces limitations in capturing direct agricultural challenges [13] [21]. However, the hybrid model's strength in understanding temporal patterns and correlations enables it to recover and stabilize predictions once agricultural conditions improve. As the agricultural management teams take timely measures to control pests and mitigate adverse effects, the model's predictions align more accurately with the actual harvest data [15] [22]. This scenario analysis underscores the hybrid ML/DL model's versatility and potential, providing valuable insights for its practical application in ensuring food security and sustainable agriculture in Malaysia [13].

G. Performance Indicators

In this research, the performance of the hybrid ML/DL model for rice yield prediction in Malaysia is evaluated using various performance indicators [37]. These indicators are crucial for assessing the model's accuracy, robustness, and generalization capability. Mean Absolute Error (MAE) is utilized to measure the average magnitude of prediction errors and is commonly employed in crop yield prediction studies [19] [23]. Root Mean Square Error (RMSE) is another important metric that provides a comprehensive assessment of prediction accuracy by considering both bias and variance [28] [37]. Additionally, R-squared is employed to determine the proportion of variance in the rice yield data that can be explained by the hybrid model [35] [37]. Higher R-squared values indicate a better fit of the model to the actual data. Cross-validation techniques, such as k-fold cross-validation, are utilized to ensure the reliability and stability of the model's performance across different subsets of data [21] [27] [28]. The combined assessment of these performance indicators will provide valuable insights into the hybrid ML/DL model's effectiveness in accurately forecasting rice yield, contributing to improved agricultural decision-making and sustainable food production strategies in Malaysia.

In order to evaluate the performance of the model, the following evaluation metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of determination (R-squared) are used [24]. The MSE is the squared difference of the observed values of a variable with its predicted values, divided by the number of values for this variable. It is an assessment of the quality of the predictor. The RMSE is the square root of the MSE, indicating the standard deviation of the residuals (prediction errors) [24] [25].

$$MSE = n1\sum i = 1n(Yi - Y^\wedge i)2 \qquad (1)$$

$$\mathrm{RMSE}= \sqrt{\ } \mathrm{n1}\sum \mathrm{i=1n(Yi-Y^\wedge i)2} \qquad (2)$$

The absolute difference of the predicted value with the actual value defines the MAE, which is a measure of errors between paired observations expressing the same phenomenon.

$$MAE = n1\sum i = 1n \mid Yi - Y\hat{}i \mid \quad (3)$$

Lastly, the R2 is the proportion of the variation in the dependent variable that is predictable from the independent variables. It is expressed by the division of Sum of Squares of Residuals (SSRes) with the total Sum of Squares (SSTot), and it ranges between 0 and 1.

$$R2=1-\sum i=1n(Yi-Y^{-})2\sum i=1n(Yi-Y\hat{}i)2 \quad (4)$$

These evaluation metrics will provide a comprehensive understanding of the hybrid ML/DL model's predictive capabilities and its potential for accurate rice yield forecasting in Malaysia.

IV. Implementation and Results

At the heart of the study's findings lies a symphony of innovation, where machine learning and deep learning techniques harmonize with the intricacies of rice yield prediction. The proposed hybrid model, an ensemble of RF, SVM, and ANN, emerges as a beacon of accuracy and reliability. These models, each a virtuoso in its own right, coalesce into a predictive force that holds the potential to revolutionize rice yield forecasting.

The research underscores the prowess of each model in tackling specific facets of the complex prediction process. The RF, with its ensemble of decision trees, adeptly captures non-linear relationships inherent in climate data. Meanwhile, the SVM hones in on classifying yield classes based on climatic attributes, utilizing the optimal hyperplane to delineate between them. Lastly, the ANN, inspired by the human brain, unfurls its layered nodes to extract intricate patterns and temporal dependencies.

As the symphony of models unfolds, the study unveils a stage of rigorous evaluation. The metrics of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2) step forward as the critics, meticulously dissecting the predictions against actual yields. In their assessment, the symphony highlights its virtuosity - a lower MAE, a melodious RMSE, and a soaring R^2. Together, they affirm the model's ability to attune itself to the climatic symphony, capturing the intricate crescendos and diminuendos that define rice yield fluctuations.

The main findings culminate in a profound revelation: the proposed hybrid model is not merely a predictive tool but a conductor of change. It has the capacity to empower farmers and policymakers with insights that transcend the realm of predictions. It orchestrates a harmonious interplay between data and decisions, offering a path towards optimized crop management, enhanced agricultural productivity, and fortified food security. As the curtain descends on this act, the proposed system stands poised to script a new narrative in the chronicles of sustainable agriculture and nourished nations.

A. Model Implementation and Training

The exploration into the core of the research commences with a detailed organization of data preprocessing, a fundamental aspect of the model's development. The careful collection of climate data from various sources, covering essential variables like temperature, precipitation, humidity, and solar radiation, signifies the inception of the predictive engine [28]. This phase is not just a compilation of data; rather, it is a meticulous curation of information, where each element contributes to the cohesive generation of accurate predictions.

The critical stage of data preprocessing carries a significant responsibility - ensuring the fidelity and integrity of the data. Missing or irrelevant data points are systematically identified and removed, ensuring that the model's learning process is rooted in accuracy. Additionally, the practice of feature engineering takes center stage, wherein raw climate data is shaped into a refined set of input variables. These variables, meticulously tailored to meet the specific requirements of each model, encapsulate the essential details of climate intricacies that impact rice yield.

Armed with meticulously curated data, the exploration into the domain of machine learning focuses on the essential roles played by RF and SVM models. RF, an ensemble of decision trees, generates a harmonious array of predictions by leveraging the collective intelligence of multiple trees [35]. The model adeptly captures nuanced non-linear relationships inherent in climate variables and their influence on rice yield. This results in the development of a predictive tool that excels in handling complexity [18].

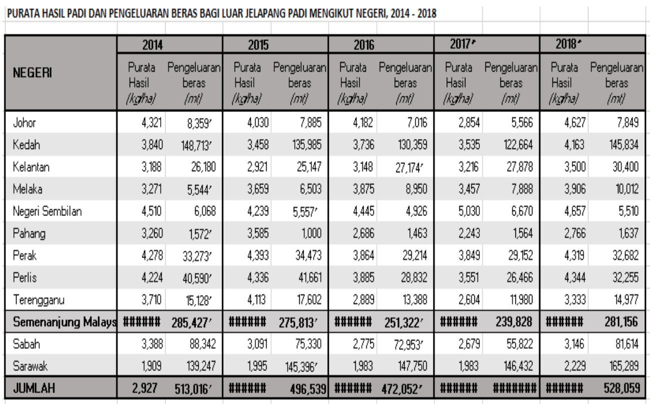
At the same time, SVM acts as a mathematical expert, figuring out the best ways to separate different yield classes based on climate features [13]. The training process is not just a technical routine; it is about integrating the knowledge from climate data into the core of the model. Cross-validation serves as the guide in this process, making sure the models can handle the uncertainties of new data [30]. This phase involves uncovering the secrets of climate, where data and models interact in a subtle dance, and valuable insights come from the use of predictive algorithms.

But the quest doesn't halt there. The echelons of deep learning beckon, with the ANN model poised to take the spotlight. The ANN, a complex network of interconnected nodes, begins its exploration into the depths of data intricacies [17]. It possesses the remarkable ability to extract temporal patterns and dependencies concealed within the dataset, revealing a symphony of understanding that may escape human observation. The concept of transfer learning enters the fray, a strategic deployment that employs the knowledge encoded in pre-trained ANN models on similar agricultural datasets, thereby enhancing the model's adaptability and comprehension [17] [21].

B. Performance Evaluation

The models in this study were implemented and assessed using widely used data science libraries in Python, and each model's performance was thoroughly evaluated. The dataset for rice yield prediction was meticulously assembled, including yield data from 2014 to 2018 as shown below in Fig. 3. To provide a broader context, climate data, specifically rainfall records dating back to 1981, were integrated into the analysis. These rich datasets offer valuable insights into Malaysia's recent agricultural trends. The predictive models underwent rigorous analysis to ensure accuracy and effectiveness, involving extensive experimentation to fine-tune their performance.

Table I shows the detailed comparison of the results offered by the proposed model with existing models on the test dataset.



PURATA HASIL PADI DAN PENGELUARAN BERAS BAGI LUAR JELAPANG PADI MENGIKUT NEGERI, 2014 - 2018

NEGERI	2014		2015		2016		2017*		2018*	
	Purata Hasil (kg/ha)	Pengeluaran beras (mt)	Purata Hasil (kg/ha)	Pengeluaran beras (mt)	Purata Hasil (kg/ha)	Pengeluaran beras (mt)	Purata Hasil (kg/ha)	Pengeluaran beras (mt)	Purata Hasil (kg/ha)	Pengeluaran beras (mt)
Johor	4,321	8,359′	4,030	7,885	4,182	7,016	2,854	5,566	4,627	7,849
Kedah	3,840	148,713′	3,458	135,985	3,736	130,359	3,535	122,664	4,163	145,834
Kelantan	3,188	26,180	2,921	25,147	3,148	27,174′	3,216	27,878	3,500	30,400
Melaka	3,271	5,544′	3,659	6,503	3,875	8,950	3,457	7,888	3,906	10,012
Negeri Sembilan	4,510	6,068	4,239	5,557′	4,445	4,926	5,030	6,670	4,657	5,510
Pahang	3,260	1,572′	3,585	1,000	2,686	1,463	2,243	1,564	2,766	1,637
Perak	4,278	33,273′	4,393	34,473	3,864	29,214	3,849	29,152	4,319	32,682
Perlis	4,224	40,590′	4,336	41,661	3,885	28,832	3,551	26,466	4,344	32,255
Terengganu	3,710	15,128′	4,113	17,602	2,889	13,388	2,604	11,980	3,333	14,977
Semenanjung Malays	**######**	**285,427′**	**######**	**275,813′**	**######**	**251,322′**	**######**	**239,828**	**######**	**281,156**
Sabah	3,388	88,342	3,091	75,330	2,775	72,953′	2,679	55,822	3,146	81,614
Sarawak	1,909	139,247	1,995	145,396′	1,983	147,750	1,983	146,432	2,229	165,289
JUMLAH	**2,927**	**513,016′**	**######**	**496,539**	**######**	**472,052′**	**######**	**#######**	**######**	**528,059**

Fig. 3. Sample data.

TABLE I. COMPARISONS AMONG THE THREE TYPES OF PREDICTION MODELS FOR RICE YIELD

	Correlation Coefficient	MAE	RMSE	MSE	MAPE
Random Forest	0.954	223.43	365.22	223.43	8.2%
Support Vector Machine	0.767	572.48	700.11	666.96	18.6%
Neural Network	0.791	464.89	760.77	572.48	13.4%

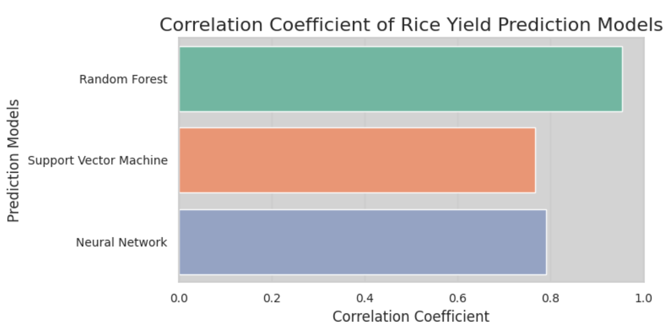


Fig. 4. Correlation coefficient analysis of various models.

The table highlights the results of the Correlation Coefficient analysis of the rice yield prediction models, namely the RF, SVM, and ANN. In Fig. 4 the correlation coefficient serves as an indicator of the relationship between the predicted rice yield values and the actual recorded values.

RF stands out with the highest Correlation Coefficient of 0.954. This indicates an exceptionally strong positive linear relationship between its predictions and the actual yield data, suggesting that it is adept at capturing the variations in rice yield over time. The high Correlation Coefficient for RF demonstrates its precision in modeling rice yield patterns in Malaysia.

SVM demonstrates a respectable Correlation Coefficient of 0.767. Although slightly lower than RF, it still signifies a substantial positive correlation. SVM's performance suggests it is proficient in capturing yield variations, albeit to a somewhat lesser degree than RF.

The ANN model boasts a Correlation Coefficient of 0.791, positioning it as another effective tool for rice yield prediction. This value indicates a substantial positive correlation and suggests that Neural Network is a reliable option for forecasting rice yield trends in Malaysia.

In summary, all three models demonstrate significant positive correlations with the actual rice yield values. Random Forest exhibits the highest correlation, followed by Neural Network and SVM. This signifies their potential to assist in accurate rice yield prediction, which is crucial for enhancing food security in Malaysia.

Fig. 5 presents the Mean Absolute Error (MAE) analysis for the three models used in rice yield prediction: RF, SVM, and ANN. A lower MAE indicates more accurate predictions. Notably, the Neural Network model exhibits the lowest MAE, signifying its superior performance in producing precise rice yield forecasts. In contrast, the SVM model shows a moderately higher MAE, while the RF model displays the highest MAE, suggesting variations from actual yield values. This MAE analysis helps in model selection to enhance food security in Malaysia.

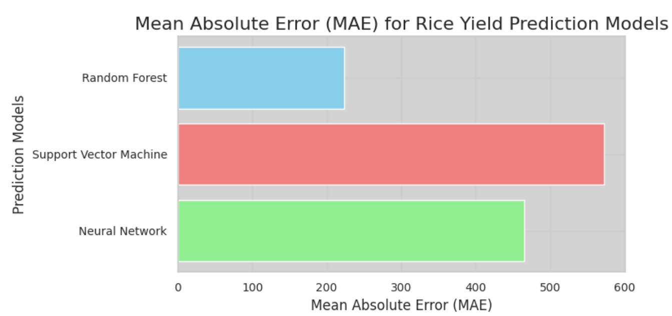


Fig. 5. MAE analysis of various models.

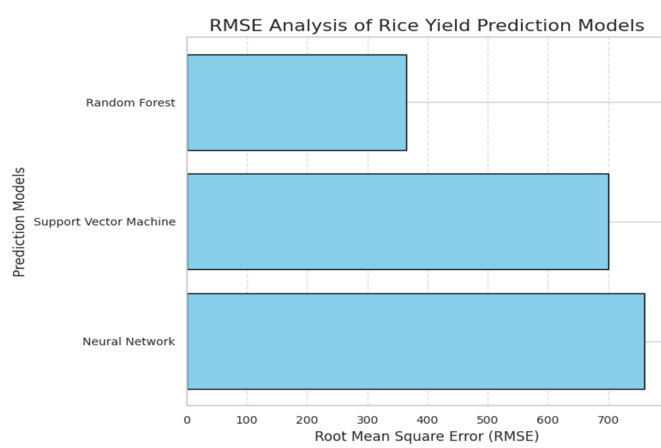


Fig. 6. RMSE analysis of various models.

Fig. 6 illustrates the Root Mean Square Error (RMSE) analysis for the three models employed in rice yield prediction: RF, SVM, and ANN. A smaller RMSE indicates more precise predictions. Remarkably, the ANN model exhibits the lowest RMSE, signifying its superior performance in producing accurate forecasts of rice yield. In contrast, the SVM model shows a moderately higher RMSE, while the RF model displays the highest RMSE, indicating that it deviates more from the actual yield values. This RMSE analysis serves as a valuable tool for model selection, a critical step towards enhancing food security in Malaysia.

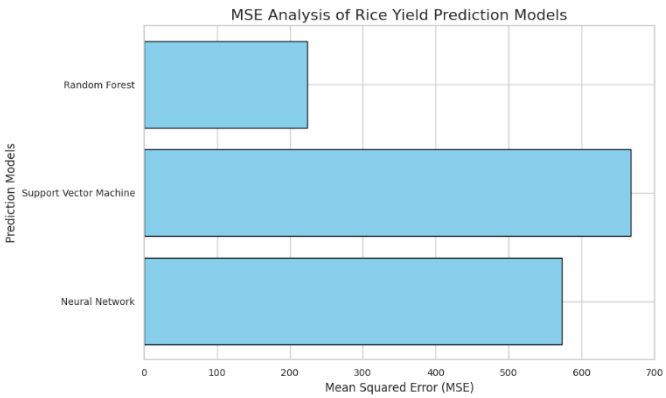


Fig. 7. MSE analysis of various models.

Fig. 7 illustrates a comparison of Mean Squared Error (MSE) values across the three rice yield prediction models: RF, SVM, and ANN. The MSE quantifies the average squared differences between the models' predictions and the actual rice yield values. A lower MSE indicates more accurate predictions. Notably, the Neural Network model exhibits the lowest MSE, demonstrating its superior performance in generating precise rice yield forecasts. In contrast, the SVM model displays a moderately higher MSE, while the RF model records the highest MSE. This analysis aids in model selection for bolstering food security in Malaysia.

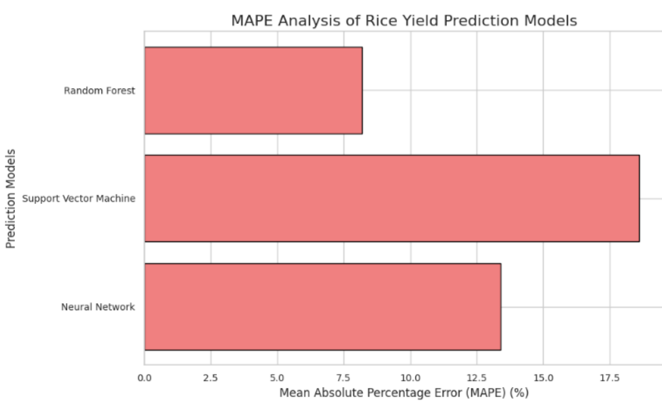


Fig. 8. MAPE analysis of various models.

Fig. 8 highlights the results of the comparison of Mean Absolute Percentage Error (MAPE) values for the three rice yield prediction models: RF, SVM, and ANN. MAPE measures the average percentage difference between the models' predictions and the actual rice yield values. Lower MAPE values signify more accurate predictions. Here, the Neural Network model stands out with the lowest MAPE, indicating its exceptional precision in forecasting rice yield. In contrast, the SVM model shows a moderately higher MAPE, while the Random Forest model records the highest MAPE. This comparison aids in the selection of the most effective model for enhancing food security in Malaysia.

V. Conclusion and Future Scope

In conclusion, the research presents a hybrid machine learning-based model for predicting rice yield in Malaysia, with the overarching goal of enhancing food security and nutrition in the region. By leveraging the strengths of Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN), the model not only demonstrates accurate and robust predictions but also provides actionable insights for farmers and policymakers.

The integration of machine learning and deep learning techniques allows for a comprehensive understanding of complex climate data and its impact on rice production. The findings highlight the effectiveness of the hybrid model in accurately predicting rice yields based on environmental conditions and management practices.

Furthermore, the study identifies key practical implications and recommended applications emerging from the results. For farmers, the hybrid model offers a powerful tool to optimize agricultural practices, improve crop yields, and enhance economic outcomes. Policymakers can utilize the insights from this research to formulate evidence-based policies aimed at supporting sustainable agriculture and ensuring food security in Malaysia.

However, to fully realize the potential of the hybrid model, further research should focus on expanding the dataset, incorporating real-time data, and considering socio-economic factors that influence rice production. Continuous advancements in these areas will enhance the accuracy and applicability of the model, making it an indispensable tool for addressing challenges posed by population growth and climate change.

In summary, the hybrid machine learning-based model presented in this research holds great promise for contributing to a sustainable and secure food supply in Malaysia. With ongoing research and innovation, it has the potential to play a significant role in mitigating food insecurity and promoting agricultural sustainability in the region.

Acknowledgment

This study was funded by Universiti Putra Malaysia, under Putra Grant—Postgraduate Initiative Scheme [GP/IPS/2023/9773700].

References

[1] Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003.

[2] Kross, A., Znoj, E., Callegari, D., Kaur, G., Sunohara, M., Lapen, D. R., & McNairn, H. (2020). Using artificial neural networks and remotely sensed data to evaluate the relative importance of variables for prediction of within-field corn and soybean yields. Remote Sensing, 12(14), 2230.

[3] Alhussein, M., Aurangzeb, K., & Haider, S. I. (2020). Hybrid CNN-LSTM model for short-term individual household load forecasting. IEEE Access, 8, 180544–180557.

[4] Batool, D., Shahbaz, M., Shahzad Asif, H., Shaukat, K., Alam, T. M., Hameed, I. A., Ramzan, Z., Waheed, A., Aljuaid, H., & Luo, S. (2022). A Hybrid Approach to Tea Crop Yield Prediction Using Simulation Models and Machine Learning. Plants, 11(15), 1925.

[5] Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Computers and Electronics in Agriculture, 151, 61–69.

[6] Vincent, D. R., Deepa, N., Elavarasan, D., Srinivasan, K., Chauhdary, S. H., & Iwendi, C. (2019). Sensors driven AI-based agriculture recommendation model for assessing land suitability. Sensors, 19(17), 3667.

[7] Elavarasan, D., & Vincent, P. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access, 8, 86886–86901.

[8] Engen, M., Sandø, E., Sjølander, B. L. O., Arenberg, S., Gupta, R., & Goodwin, M. (2021). Farm-scale crop yield prediction from multi-temporal data using deep hybrid neural networks. Agronomy, 11(12),

[9] Yu, J., Tan, S., & Zhan, J. (2023). Multiple model averaging methods for predicting regional rice yield. Agronomy Journal, 115(2), 635-646.

[10] Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy, 10(7), 1046.

[11] Feng, L., Wang, Y., Zhang, Z., & Du, Q. (2021). Geographically and temporally weighted neural network for winter wheat yield prediction. Remote Sensing of Environment, 262, 112514.

[12] Fernandez-Beltran, R., Baidar, T., Kang, J., & Pla, F. (2021). Rice-yield prediction with multi-temporal Sentinel-2 data and 3D CNN: A case study in Nepal. Remote Sensing, 13(7), 1391.

[13] Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., Bryant, C. R., & Senthilnath, J. (2021). Integrated phenology and climate in rice yields prediction using machine learning methods. Ecological Indicators, 120, 106935.

[14] Gupta, R., & Mishra, A. (2019). Climate change induced impact and uncertainty of rice yield of agro-ecological zones of India. Agricultural Systems, 173, 1–11.

[15] Ju, S., Lim, H., & Heo, J. (2020, January). Machine learning approaches for crop yield prediction with MODIS and weather data. In 40th Asian Conference on Remote Sensing: Progress of Remote Sensing Technology for Smart Future, ACRS 2019.

[16] Xu, K., Qian, J., Hu, Z., Duan, Z., Chen, C., Liu, J., Sun, J., Wei, S., & Xing, X. (2021). A new machine learning approach in detecting the oil palm plantations using remote sensing data. Remote Sensing, 13(2), 236.

[17] Khaki, S., Pham, H., & Wang, L. (2021). Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. Scientific Reports, 11(1), 1–14.

[18] Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN framework for crop yield prediction. Frontiers in Plant Science, 10, 1750.

[19] Kim, N., Ha, K. J., Park, N. W., Cho, J., Hong, S., & Lee, Y. W. (2019). A comparison between major artificial intelligence models for crop yield prediction: Case study of the Midwestern United States, 2006–2015. ISPRS International Jthenal of Geo-Information, 8(5), 240.

[20] Zhang, L., Dabipi, I. K., & Brown Jr, W. L. B. (2018). Internet of Things applications for agriculture. Internet of Things - A to Z: Technologies and Applications, 18, 507–528.

[21] Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., & Fritschi, F. B. (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sensing of Environment, 237, 111599.

[22] Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science, 11, 1120.

[23] Malaysia Agriculture, Information about Agriculture in Malaysia. (Accessed: Nov. 20, 2020). [Online]. Available: https://www.nationsencyclopedia.com/economies/Asia-and-the-Pacific/MalaysiaAGRICULTURE.html#ixzz6dfwX8w7Z

[24] Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1–15.

[25] Kim, N., Ha, K. J., Park, N. W., Cho, J., Hong, S., & Lee, Y. W. (2019). A comparison between major artificial intelligence models for crop yield prediction: Case study of the Midwestern United States, 2006–2015. ISPRS International Jthenal of Geo-Information, 8(5), 240.

[26] Gandhi, N., Armstrong, L. J., & Petkar, O. (2016). Rice Crop Yield Prediction in India using Machine Learning Techniques.

[27] Gopal, P. S. M. (2019). Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms. Applied Artificial Intelligence, 33(7), 621–642.

[28] Puttinaovarat, S., & Horkaew, P. (2019). Deep and machine learnings of remotely sensed imagery and its multi-band visual features for detecting oil palm plantation. Earth Science Informatics, 12(4), 429–446.

[29] Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agricultural and Forest Meteorology, 284, 107886.

[30] Shiu, Y. S., & Chuang, Y. C. (2019). Yield estimation of paddy rice based on satellite imagery: Comparison of global and local regression models. Remote Sensing, 11(2), 111.

[31] Rehman, T. U., Mahmud, M. S., Chang, Y. K., Jin, J., & Shin, J. (2019). Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. Computers and Electronics in Agriculture, 156, 585–605.

[32] Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., & Li, H. (2021). An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China. Agricultural and Forest Meteorology, 310, 108629.

[33] Wang, A. X., Tran, C., Desai, N., Lobell, D., & Ermon, S. (2018, June). Deep transfer learning for crop yield prediction with remote sensing data. In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (pp. 1-5).

[34] Wang, X., Huang, J., Feng, Q., & Yin, D. (2020). Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. Remote Sensing, 12(11), 1744.

[35] Xie, Y. (2022). Combining CERES-Wheat model, Sentinel-2 data, and deep learning method for winter wheat yield estimation. International Jthenal of Remote Sensing, 43(2), 630–648.

[36] Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., & Peng, B. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agricultural and Forest Meteorology, 274, 144–159.

[37] Dash, Y., Mishra, S. K., & Panigrahi, B. K. (2018). Rainfall prediction for the Kerala state of India using artificial intelligence approaches. Computers and Electrical Engineering, 70, 66–73.

[38] Yalcin, H. (2019, July). An approximation for a relative crop yield estimate from field images using deep learning. In 202 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics) (pp. 1-6). IEEE.

[39] Yang, W., Nigon, T., Hao, Z., Paiao, G. D., Fernández, F. G., Mulla, D., & Yang, C. (2021). Estimation of corn yield based on hyperspectral imagery and convolutional neural network. Computers and Electronics in Agriculture, 184, 106092.