

RICE YIELD PREDICTION USING MACHINE LEARNING

Thuy-Vi Thi Ha¹, Thanh-Nghi Do^{2*}, Bao-An Nguyen³

Abstract – *This paper proposes a novel approach for predicting rice yield using rice field images. This approach utilizes the ability of Vision Transformer (ViT) architecture to extract meaningful features from field images for rice yield prediction. The model was first trained with a classification task. The standard Vision Transformer model is modified by replacing the classification layer with a custom regression layer designed to predict rice yield. This modified Vision Transformer model is then trained on field images with corresponding yield data. Various regression models, such as random forests (RF), support vector regressors (SVR), and multi-layer perceptrons (MLP), were employed to find the best regression model for rice yield prediction. Over 11,000 digital images were collected during the ripening stage of rice plants in An Giang Province and Tra Vinh Province (Vietnam), with the rough grain yield recorded after harvest in these areas ranging from 5 to 12 t ha⁻¹. The experimental results indicate that Vision Transformer – Random forests model achieved the lowest mean absolute error is 75.96.*

Keywords: *multi-layer perceptron (MLP), random forests (RF), rice yield prediction, support vector machines (SVM), Vision Transformer (ViT).*

I. INTRODUCTION

Rice is very important to many countries around the world, especially countries in Asia. It is the main source of food for more than half of the world's population [1]. Rice is essential for

food security, addressing the increasing food demand due to population growth, and maintaining social stability [2]. Accurate predictions of rice yields will help optimize agricultural planning, improve food security, and support decision-making for farmers and agricultural managers.

Researchers extracted features from images to estimate rice yield based on machine learning models [3–6]. The spectral reflectance characteristics, vegetation indices (VIs), canopy height, canopy cover, normalized difference vegetation index (NDVI), leaf area index (LAI), color index (CI), and near-infrared data are collected from multispectral and hyperspectral images of satellite remote sensing. These features are utilized to predict rice yield employing conventional models such as linear regression (LR), support vector machine (SVM), random forest (RF), decision tree (DT), and K-nearest neighbor (KNN) [3–5]. In addition, advanced machine learning models such as long short-term memory (LSTM), bidirectional LSTM (Bi-LSTM), Gaussian process regression (GPR), fuzzy inference system (FIS), FIS adaptive neural system (ANFIS), M5 model tree (M5 Tree), support vector regression (SVR), RF, and powerful ensemble techniques based on Bayesian model averaging (BMA) are also used to predict rice yield as a new improvement in Sarkar's study [6]. However, feature extraction methods from multispectral and hyperspectral images using satellite remote sensing technology still have many limitations. These include poor-quality images due to distance, cloud cover, and weather factors during collection. Features are extracted manually and depend on the decisions of experts in the field. Thus, the process of collecting and processing data requires significant effort and money.

Scientists have increasingly adopted deep learning neural networks to autonomously learn

^{1,3}Tra Vinh University, Vietnam

²Can Tho University, Vietnam

*Corresponding author: dtngghi@cit.ctu.edu.vn

Received date: 28th June 2024; Revised date: 24th July 2024; Accepted date: 02nd August 2024

features from input data without depending on experts [7–12]. Hybrid models combining convolutional neural networks (CNN) with LR, RF, and DT have been used to predict rice yield based on rice panicle images [7]. Crop yield is related to the number of panicles per square meter, grains per panicle, and grain size [8–11]. Therefore, counting panicles is also an appropriate method for predicting rice yield. Models such as You Only Look Once (YOLO) and Faster Region-based Convolutional Neural Network (Faster R-CNN) have been explored for detecting and counting rice panicles [12]. However, these methods are only suitable for small-scale sampling areas due to the time-consuming process of isolating grains and the heterogeneous crop density across different regions. CNN models using Red-Green-Blue (RGB) images also show promise in rice yield prediction [13–15] and need many further improvements.

The study proposes two new approaches to improve the performance of the rice yield prediction model. Firstly, the dataset of rice field images is collected from diverse devices such as unmanned aerial vehicles (UAVs), digital cameras, and smartphone cameras in Tra Vinh and An Giang Provinces. Secondly, ViT’s powerful ability is leveraged to extract useful image features and then use conventional models such as RF, SVR, or multi-layer perceptrons (MLP) to improve prediction model accuracy.

The remainder of this paper is organized as follows. Section 2 presents the experimental methods. Section 3 shows the experimental results before conclusions and future works presented in Section 4.

II. PREDICTING RICE YIELD WITH VISION TRANSFORMER AND REGRESSORS

This investigation aims to propose a machine learning flow to accurately predict rice yield from images (as shown in Figure 1). The main idea consists of two tasks: 1) training a Vision Transformer (ViT) model on field images to predict rice yield; 2) using the ViT model [16] to extract

visual features from field images, followed by training machine learning models such as RF [17], SVM [18], or MLP [19] on the extracted features to predict rice yield.

A. Training vision transformer

The ViT is a type of neural network architecture designed specifically for computer vision tasks [16]. It adapts the transformer model, which has been highly successful in natural language processing (NLP), to the domain of image analysis. The ViT architecture includes main blocks such as image tokenization, linear projection, position embedding, transformer encoder, classification token, and output layer. The ViT has advantages: global context understanding, scalability, reduced inductive bias (do not impose a local connectivity pattern – convolutions), and typically transfer learning.

Pre-trained ViT can be fine-tuned on smaller datasets, leveraging the knowledge acquired from large-scale pre-training on ImageNet [20]. This is exactly the case for rice yield prediction in this research. The study proposes to replace the MLP head (unfilled block in Figure 1) with new layers for the regression problem of predicting rice yield from field images as (*).

$$Dense(1024, activation = "relu") \Rightarrow Dropout(0.33) \Rightarrow Dense(1) \quad (*)$$

Then, the new ViT trains the regression model from field images to predict rice yield.

B. Training regressors

The ViT model learned from field images for extracting visual features was utilized. Thus the research proposes to train regression models with learning algorithms of RF, SVM, or MLP, to predict rice yield.

An RF regressor learns an ensemble of maximum depth decision trees from bootstrap samples, for each node in the tree, the best splitting is chosen from a random subset of features. The final prediction of the RF regressor is obtained by averaging the predictions from all individual trees.

SVR tries to find an optimal hyperplane in a high-dimensional space that best represents the relationship between the input variables and the target variable. SVR aims to minimize the error between the actual and predicted values while maintaining a margin of tolerance. SVR is effective in handling non-linear relationships through the use of kernel functions.

An MLP consists of multiple layers of nodes (neurons), including an input layer, one or more hidden layers, and an output layer. Each neuron in the network is connected to neurons in the adjacent layers, and each connection has an associated weight. During training, the network learns to adjust these weights to minimize the difference between the predicted and actual target values. The study proposes an MLP architecture with a hidden layer for predicting rice yield from field images as (**).

$Dense(1024, activation = "relu") \Rightarrow$
 $Dropout(0.33) \Rightarrow Dense(1) \quad (**)$

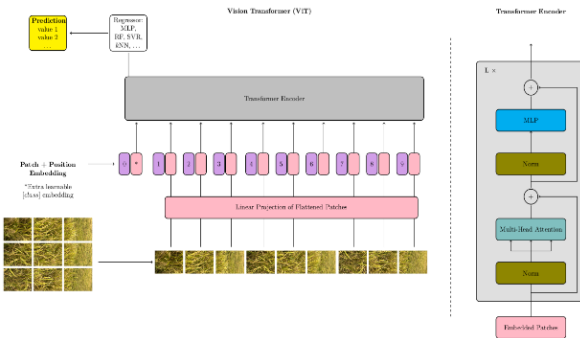


Fig. 1: Vision Transformer (ViT) and regressors for predicting rice yield

III. EXPERIMENTAL RESULTS

The assessment of the proposed ViT and regressors for predicting rice yield is of interest to us. Consequently, there is a need to evaluate its performance concerning prediction results.

A. Software programs

The research implemented ViT, Vision Transformer – Support vector regression (ViT-SVR),

Vision Transformer – Multi-layer perceptrons ViT-MLP, and Vision Transformer – Random forests (ViT-RF) in Python using libraries Scikit-learn [21], and Tensorflow [22].

All experiments used a machine Linux Fedora 32, Intel(R) Core i7-4790 CPU, 3.6 GHz, 4 cores and 32 GB main memory, and the Nvidia Gigabyte GeForce RTX 2080Ti 11GB GDDR6, 4352 CUDA cores.

B. Dataset

Rice canopy images were collected from 47 plots in An Giang and Tra Vinh Provinces. The dataset of rice field images includes 11,736 images collected from various devices such as UAVs, digital cameras, and smartphone cameras during the ripening stage. The images are diverse and captured from varying angles, distances, and resolutions. The number of images from each device type is irregular. Additionally, the corresponding approximate grain yield was recorded, which ranged from 5 to 12 t ha⁻¹ as shown in Table 1.

To train and validate the model, the dataset is randomly divided into a training set with 8,000 images and a testing set with 3,736 images.

C. Tuning parameters




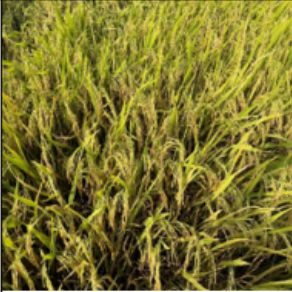








The ViT regression model is trained with the parameter setting: $batch_size = 4$, $epochs = 15$, optimizers. Adam ($learning_rate = 0.001$).

The SVR algorithm learns the regression model using a radial basis function (RBF) with $gamma = 0.0001$, $epsilon = 0.1$, and $C = 1000000$.

The MLP algorithm learns the regression model with $max_iter = 100$, $batch_size = "auto"$, $learning_rate = "constant"$, $learning_rate_init = 0.001$, $power_t = 0.5$.

The RF algorithm trains 50 trees with $n_estimators = 50$, $criterion = mae$, $min_samples_split = 5$, $min_samples_leaf = 5$, $max_features = 20$.

Table 1: Some pictures of rice fields along with actual yields

Rice yield (kg/1000 m ²)	Pictures of rice fields		
900.133			
1020.500			
670.000			
770.986			

D. Prediction results

To evaluate the prediction, the mean absolute error (MAE) common measure is used to evaluate the error of a predictive model. It represents the average absolute difference between the predicted values and the actual values. The lower the MAE is, the more accurate the model’s predictions are.

The formula for calculating MAE is given by Equation (1).

where:

n is the number of predictions,

y_i is the actual value,

y^i is the predicted value.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

Table 2: Prediction results

No.	Method	Prediction error (MAE)
1	ViT	84.18
2	ViT-SVR	79.31
3	ViT-MLP	77.8
4	ViT-RF	75.96

From the results in Table 2, it is clear that the use of ViT with machine learning algorithms improves prediction accuracy. The standalone ViT model has an MAE of 84.18. The ViT-SVR decreases the MAE to 79.31. The ViT-MLP further reduces the MAE to 77.80. The best performance is achieved with ViT-RF, which attains the lowest MAE of 75.96. These results demonstrate that the ViT-RF model provides the most accurate yield predictions among the evaluated methods as shown in Figure 2.

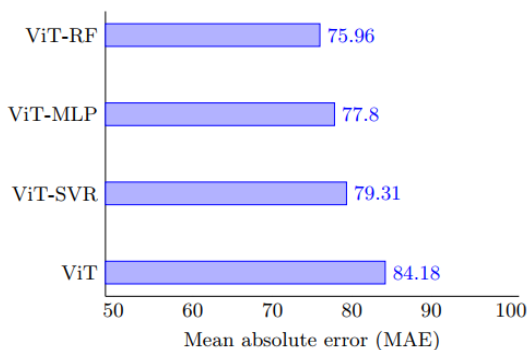


Fig. 2: Prediction results (MAE)

IV. CONCLUSION AND FUTURE WORKS

This paper proposed a novel approach for predicting rice yield using field images. The method leverages the ViT model's powerful feature extraction capabilities and combines them with traditional machine learning models to enhance prediction accuracy. First, a ViT model for regression task is trained by replacing the classification MLP head with regression MLP layers. The trained ViT model is then used to extract features from the field images. These features are subsequently used to predict rice yield based on machine learning models such as RF, SVR, and MLP.

The study employed devices such as UAVs, digital cameras, and smartphones to collect high-resolution images. The dataset includes over 11,000 images captured during the ripening stage in the An Giang and Tra Vinh Provinces. While previous studies primarily used images obtained

from satellite remote sensing or UAV technology integrated with spectral sensors, this approach incorporates these diverse imaging sources. Rough grain yields recorded after harvest in these areas range from 5 to 12 t ha⁻¹. Experimental results demonstrate that the ViT-RF model achieves the lowest MAE of 75.96 kg per 1000 m².

In the near future, the research team will develop an incremental learning algorithm to further enhance the model's performance across different stages of rice growth. This approach will allow the model to continuously learn and adapt from new data, accommodating changes and variations in the rice fields throughout the growing season. This incremental learning process will enable the model to update its knowledge without retraining from scratch, improving efficiency and maintaining high prediction accuracy over time. The research demonstrates the feasibility of the ViT model in agriculture.

REFERENCES

- [1] Xu T, Wang F, Yi Q, Xie L, Yao X. A bibliometric and visualized analysis of research progress and trends in rice remote sensing over the past 42 years (1980–2021). *Remote Sensing*. 2022;14(15): 3607. <https://doi.org/10.3390/rs14153607>.
- [2] Matsumura K, Hijmans RJ, Chemin Y, Elvidge CD, Sugimoto K, Wu W, et al. Mapping the global supply and demand structure of rice. *Sustainability Science*. 2009;4(2): 301–313. <https://doi.org/10.1007/s11625-009-0077-1>.
- [3] Wan L, Cen H, Zhu J, Zhang J, Zhu Y, Sun D, et al. Grain yield prediction of rice using multi-temporal UAV-based RGB and multispectral images and model transfer – a case study of small farmlands in the South of China. *Agricultural and Forest Meteorology*. 2020;291: 108096. <https://doi.org/10.1016/j.agrformet.2020.108096>.
- [4] Zhou X, Zheng HB, Xu XQ, He JY, Ge XK, Yao X, et al. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multi-spectral and digital imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2017;130: 246–255. <https://doi.org/10.1016/j.isprsjprs.2017.05.003>.
- [5] Son NT, Chen CF, Chen CR, Guo HY, Cheng YS, Chen SL, et al. Machine learning approaches for rice crop yield predictions using time-series satellite data in Taiwan. *International Journal of Remote Sensing*. 2020;41(20): 7868–7888. <https://doi.org/10.1080/01431161.2020.1766148>.

- [6] Sarkar TK, Roy DK, Kang YS, Jun SR, Park JW, Ryu CS. Ensemble of machine learning algorithms for rice grain yield prediction using UAV-based remote sensing. *Journal of Biosystems Engineering*. 2024;49(1): 1–19. <https://doi.org/10.1007/s42853-023-00209-6>.
- [7] Pankaj, Kumar B, Bharti PK, Vishnoi VK, Kumar K, Mohan S, et al. Paddy yield prediction based on 2D images of rice panicles using regression techniques. *The Visual Computer*. 2024;40(6): 4457–4471. <https://doi.org/10.1007/s00371-023-03092-6>.
- [8] Slafer GA, Savin R, Sadras VO. Coarse and fine regulation of wheat yield components in response to genotype and environment. *Field Crops Research*. 2014;157: 71–83. <https://doi.org/10.1016/j.fcr.2013.12.004>.
- [9] Lu H, Cao Z, Xiao Y, Fang Z, Zhu Y, Xian K. Fine-grained maize tassel trait characterization with multi-view representations. *Computers and Electronics in Agriculture*. 2015;118: 143–158. <https://doi.org/10.1016/j.compag.2015.08.027>.
- [10] Ferrante A, Cartelle J, Savin R, Slafer GA. Yield determination, interplay between major components and yield stability in a traditional and a contemporary wheat across a wide range of environments. *Field Crops Research*. 2017;203: 114–127. <https://doi.org/10.1016/j.fcr.2016.12.028>.
- [11] Jin X, Liu S, Baret F, Hemerlé M, Comar A. Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery. *Remote Sensing of Environment*. 2017;198: 105–114. <https://doi.org/10.1016/j.rse.2017.06.007>.
- [12] Wang X, Yang W, Xiong L, Huang C, Liang X, Chen G, et al. Field rice panicle detection and counting based on deep learning. *Frontiers in Plant Science*. 2022;13: 966495. <https://doi.org/10.3389/fpls.2022.966495>.
- [13] Yang Q, Shi L, Han J, Zha Y, Zhu P. Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. *Field Crops Research*. 2019;235: 142–153. <https://doi.org/10.1016/j.fcr.2019.02.022>.
- [14] Tanaka Y, Watanabe T, Katsura K, Tsujimoto Y, Takai T, Tanaka TST, et al. Deep learning enables instant and versatile estimation of rice yield using ground-based RGB images. *Plant Phenomics*. 2023;5: 0073. <https://doi.org/10.34133/plantphenomics.0073>.
- [15] Bellis ES, Hashem AA, Causey JL, Runkle BRK, Moreno-García B, Burns BW, et al. Detecting intra-field variation in rice yield with unmanned aerial vehicle imagery and deep learning. *Frontiers in Plant Science*. 2022;13: 716506. <https://doi.org/10.3389/fpls.2022.716506>.
- [16] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria*. 3–7 May 2021; Vienna, Austria. ICLR; 2021. <https://doi.org/10.48550/arXiv.2010.11929>.
- [17] Breiman L. Random forests. *Machine Learning*. 2001;45: 5–32. <https://doi.org/10.1023/A:1010933404324>.
- [18] Vapnik VN. *The nature of statistical learning theory*. New York: Springer; 1995.
- [19] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 1998;86(11): 2278–2324. <https://doi.org/10.1109/5.726791>.
- [20] Deng J, Berg A C, Li K, Fei-Fei L. What does classifying more than 10,000 image categories tell us? In: Daniilidis K, Maragos P, Paragios N (eds.). *Computer Vision – ECCV 2010 – 11th European Conference on Computer Vision*. 5–11 September 2010; Heraklion, Crete, Greece. Berlin: Springer Berlin Heidelberg; 2010. p.71–84. https://doi.org/10.1007/978-3-642-15555-0_6.
- [21] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011;12: 2825–2830. <https://doi.org/10.5555/1953048.2078195>.
- [22] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv [Preprint]* 2016. Version 2. <https://doi.org/10.48550/arXiv.1603.04467>.

