Rice Crop Field Classification based on Satellite Data

Oshmita Sarkar[*], Ruchira Saha[*], Devashree Baruah[*] , Sanchari Ray[*]

[*] Kalinga Institute of Industrial Technology, Bhubaneswar, India

Abstract - **The project aims to develop an accurate and efficient model for rice crop identification that can be used to improve agricultural productivity and, ultimately, alleviate hunger. The data set was gathered from the European Copernicus (Sentinel-1 and Sentinel-2) program and NASA's Landsat program. This project used data from the An Giang province in the Mekong Delta in Vietnam. The rice crop yield data was collected for the period of late-2021 to mid-2022 over the Chau Phu, Chau Thanh and Thoai Son districts. The data set contains information about the location of rice crops and non-rice crops. Band values such as VV[gamma-nought values of the signal transmitted with vertical polarization and received with vertical polarization with radiometric terrain correction applied] and VH[gamma-nought values of signal transmitted with vertical polarization and received with horizontal polarization with radiometric terrain correction applied] are obtained using the data set. These bands help in distinguishing between the rice and the non-rice crops. Commonly used machine learning algorithms such as gradient boosting classifier, label propagation, random forest classifier, KNN[K - Nearest Neighbour] and Gaussian process classifier were used to predict the presence of rice crops. Better results were obtained for the Gaussian process classifier with an f1-score of 0.90.**

Index Terms- **crop identification, machine learning, Sentinel-1 and Sentinel-2 satellite, f1-score, VV and VH bands.**

I. Introduction

1) The growing demand for precise and effective agricultural monitoring has made crop identification an important research topic in recent years. Crop identification is even more important in Vietnam, where hunger is still a major problem since it can help end hunger. Although Vietnam has a robust agricultural sector, the country still has a large population living in poverty and insecure access to food. Because rice is the main source of food for the majority of Vietnamese people, it is important to monitor this crop. The identification of rice crops using data from the An Giang province in the Mekong Delta in Vietnam is the main topic of this research article.

The growth and health of vegetation are evaluated using vegetation indices like the Normalized Difference Vegetation Index (NDVI) and the Ratio Vegetation Index (RVI). These vegetation indices are crucial resources for differentiating between various crop kinds and evaluating the productivity and health of those crops in the context of crop identification. We can learn a lot about the growth patterns of crops in Vietnam by examining their NDVI and RVI values, and we can also spot any potential problems or pressures that might be influencing their output. The eradication of hunger in Vietnam can then be achieved by using this knowledge to create targeted interventions and enhance farming methods.

II. BASIC CONCEPTS/ RELATED WORK

This project tries to forecast the Rice crop field ML classification based on satellite data.

In this project, we have used sklearn library and we have imported libraries like odc, ipyleaflet, matplotlib, seaborn,numpy, pandas, pystac, pystac_client, rich.table and requests.

Some machine learning methods like gradient descent, ada boosting, decision tree classifier, K-nearest classifier, BernoulliNB, Logistic regression CV etc are used to compare the prediction model. The

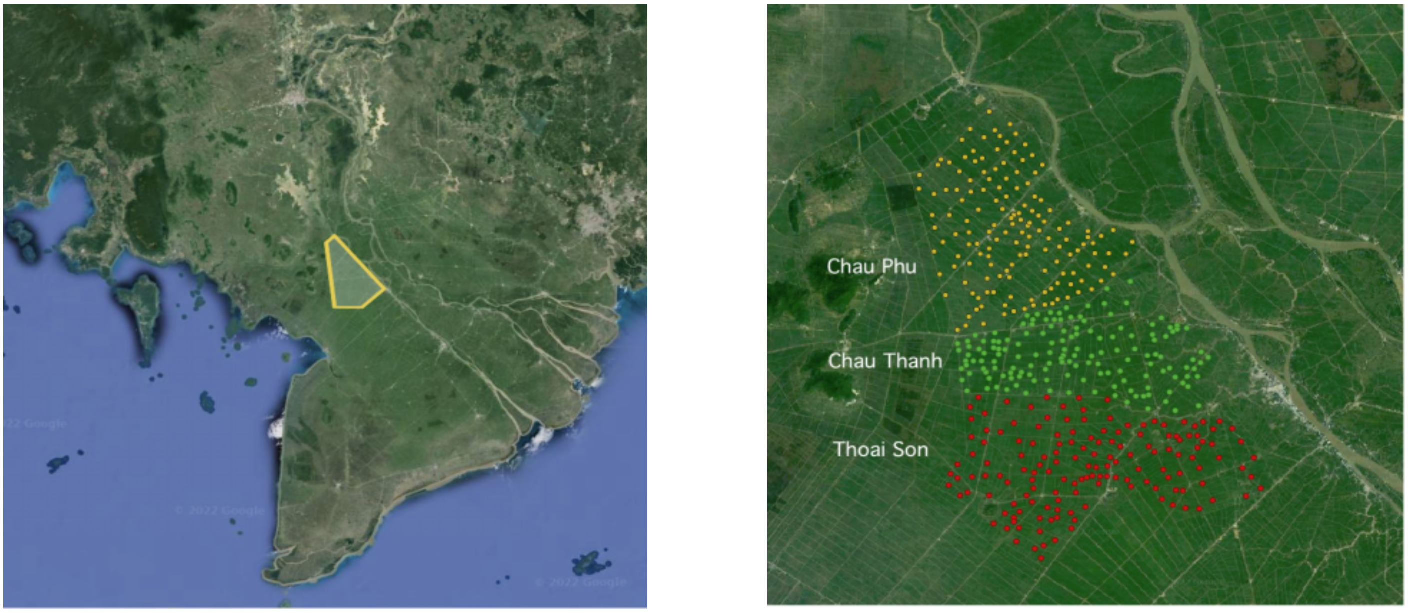


Figure: *For the three central districts of Chau Phu, Chau Thanh, and Thoai Son in the Vietnamese province of An Giang, rice crop statistics are available.*

Parameters such as EVI, NDVI, RVI, and SAVI are taken into consideration for the rice crop field classification. Among the ML algorithms we have used, Gaussian Process Classifier gives the best result with an accuracy of 90%.

Extracting the trends such as NDVI, EVI, RVI and SAVI which were present in the meteorological data was collected from NASA’s Landsat program. Following data collection, image processing, NDVI calculation, and then application of the ML technique i.e. Extreme Gradient Boosting for crop production forecasting. In Vietnam, the model was used to forecast the rice crop. The factors include the highest and lowest temperatures, vapor pressure, the amount of precipitation, etc. It involves the following steps:

Calculating NDVI:

The Normalised Difference Vegetation Index (NDVI) is a visual indicator created using data from remote sensing.
From a space station, it is utilized to assess the green vegetation in general.

.

$$\text{NDVI} = \frac{(\text{NIR - Red})}{(\text{NIR + Red})}$$

Near-InfraRed is referred to as NIR. It is possible to see the color red. The NDVI value ranges from -1 to +1. It was once used to gauge how densely vegetated a plot of land was.

Calculating SAVI:

The Soil Adjusted Vegetation Index (SAVI), which takes into consideration the impact of soil brightness on the NDVI, is employed in areas with little vegetation cover. To mitigate the effects of soil brightness, the SAVI was developed. A soil adjustment factor L was added to the NDVI equation to account for the impacts of soil noise (soil color, soil moisture, soil variability between regions, etc.).

$$SAVI = \frac{(NIR - RED)}{(NIR + RED + L)} * (1 + L)$$

Calculating EVI:

An indicator of the amount of vegetation in a certain area is the enhanced vegetation index (EVI). created to boost the vegetation signal (LAI) and NDVI in areas with high leaf area indices. Using the blue reflection section, the EVI minimizes air factors such as aerosol scattering and corrects background soil indications

$$EVI = G \times \frac{(NIR - Red)}{NIR + C1 \times Red - C2 \times Blue + L}$$

Where L stands for canopy backdrop adjustment, NIR/red/blue is the auto-corrected surface reflectance and C1, C2 are the aerosol resistance factor coefficients. EVI has a value between -1 and 1. The EVI is very helpful for evaluating crop phenology mapping, crop growth patterns, and agricultural production estimates, as well as for identifying desired land use and land cover aspects.

Gaussian Process Classifier

The GPC is especially helpful for classification jobs where the distinctions between various classes are ambiguous or ill-defined. This is so because the GPC uses a Gaussian distribution to characterize each class, allowing for a classification boundary that is more forgiving. We must first acquire satellite photos of the rice fields in order to implement a GPC for the classification of rice crop fields. The features that are important for classification, like vegetation indices or texture properties, can then be extracted from these photos. Once we have our feature set, we can train the GPC model using a training set of labeled data. The classification of fresh, unused satellite photos of rice fields can then be done using the GPC model.

Satellite data typically involves high-dimensional, complex data that can exhibit non-linear relationships between features and labels. In this context, GPC can be a particularly effective tool because it allows for flexible modeling of complex relationships in the data.

GPC works by modeling the distribution of possible functions that could fit the data, based on prior knowledge about the data and the observed input-output pairs. This distribution is represented by a Gaussian process, which is essentially a collection of random variables that describe the possible functions that could fit the data.

In the case of satellite data, GPC can be effective because it can capture complex relationships between features such as vegetation indices, temperature, humidity, etc., and the corresponding output labels such as land cover types, precipitation levels, etc. This is important because traditional linear models may not be able to capture the full complexity of these relationships, leading to underfitting or poor performance.

Moreover, GPC has several other advantages that make it well-suited for satellite data. For example, it can handle missing data, can work with noisy or uncertain labels, and can adaptively choose hyperparameters that maximize the predictive accuracy of the model.

Sentinel-1 Data

The C-band synthetic-aperture radar equipment aboard the Sentinel-1 spacecraft collects data in all weather conditions, day or night. This equipment has a 410 km sweep and spatial resolution down to 5 m. The satellite travels in an almost polar (98.18° inclination), sun-synchronous orbit.

Sentinel-2 Data

Collecting Sentinel-2 data, including satellite photos, weather, rainfall, and soil type information, and then combining it in a structured way before cleaning the data. Cleaning up data is essential because it increases overall productivity by enhancing data quality by removing inaccurate, unsuitable, and incomplete data.

III. LITERATURE REVIEW

Crop growth is essential for determining the crop's yield. To determine crop development and growth, process-based models make use of multiple crops, environmental, and other management practices. However, despite having important statistics, they do not fully represent the factors that contribute to crop yield decline. Through the use of satellite images, remote sensing gathers the most recent information on crops. The satellite photos are available for free under open information strategies and are error-free. However, because crop yield can only be observed indirectly through satellite data, statistical models must be used to predict crop yield from satellite observations. The outputs from the previous approaches as well as meteorological variables are used as predictors in statistical models.

In a 2017 study by Anh et al., rice crop regions in Vietnam's Mekong Delta were accurately mapped with 92% accuracy using Landsat 8 imagery.Nguyen et al. (2019) employed a multi-level SVM classifier to categorize rice crops in the Vietnam Red River Delta region with an overall accuracy of 94.67% in another study.Based on satellite photos, it has been demonstrated that the application of machine learning algorithms such as Random Forest, Support Vector Machines, and Maximum Likelihood can successfully detect and categorize rice crop areas.Open-source satellite data, including those from Landsat and Sentinel-2, has given academics a multitude of tools with which to conduct their research.

Findings: Studies have found that rice can be classified with high accuracy rates of between 80% and 97%. According to several studies, rice yield and spectral indexes like the NDVI (Normalized Difference Vegetation Index) are correlated.

Limitations: The temporal and spatial resolution of satellite imagery, which can lead to inaccuracies in crop classification, is one limitation of prior studies. Additionally, for accurate ground truth data collection, human intervention may be necessary.

IV. PROPOSED WORK

1) Data Preprocessing:

Data Cleaning has been done in the project as the data over a time frame has been collected. The data frame comprises the Latitude and Longitude of the given satellite images. The latitude and longitude are specified in the Sentinel-1 RTC Data. From the given latitudes and longitudes, we have calculated the vertical-horizontal polarization and vertical-vertical polarization indices. They are further merged into the data frame. In the ground truth values of the training data set, the images are labeled as rice and non-rice fields. We proposed generating a bounding box around the images and based on that we determined the vv and vh indices.
We have pulled out the API key from the Microsoft planetary computer and have calculated the Radar Vegetation Index (RVI) on the given vv and vh values.

2) Resolution Selection

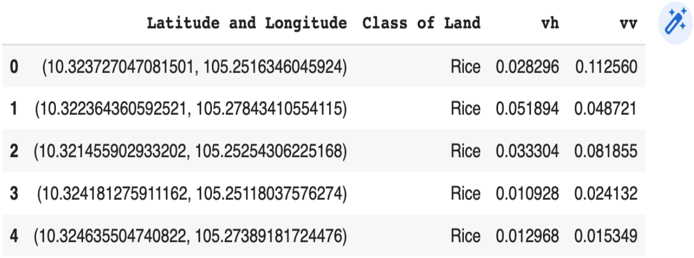
The resolution that was first specified was 10 meters/pixel.
The scale on which the images are examined is the 11132nd fraction of the resolution.
For the given project, we have stuck to the EPSG:4326 coordinate resolution scale. We have used this CRS since the account is of Vietnamese crop fields and Vietnam is placed in the GPS coordinates, consisting of a latitude of 14.0583° N and a longitude of 108.2772° E. Hence Vietnam lies in the Eastern and Northern Hemispheres.
There is a detailed account of this particular CRS as per epsg.io

WGS 84 -- WGS84 - World Geodetic System 1984, used in GPS

Attributes
Unit: degree (supplier to define representation)
Geodetic CRS: WGS 84
Datum: World Geodetic System 1984 ensemble
Data source: EPSG
Information source: EPSG. See 3D CRS for an original information source.
Revision date: 2020-03-14
Scope: Horizontal component of the 3D system.

Area of use: World.
Coordinate system: Ellipsoidal 2D CS. Axes: latitude, longitude. Orientations: north, east. UoM: degree

3) Sample Dataset:



	Latitude and Longitude	Class of Land	vh	vv
0	(10.323727047081501, 105.2516346045924)	Rice	0.028296	0.112560
1	(10.322364360592521, 105.27843410554115)	Rice	0.051894	0.048721
2	(10.321455902933202, 105.25254306225168)	Rice	0.033304	0.081855
3	(10.324181275911162, 105.25118037576274)	Rice	0.010928	0.024132
4	(10.324635504740822, 105.27389181724476)	Rice	0.012968	0.015349

4) Models and Hyperparameter tuning:

VH and VV bands were computed and the feature matrix was prepared with VH and VV band values as the columns. Our target variable is “Class of Land” (Rice or non-rice in this case). The data was then split into training and test sets by keeping 30% of the data for testing. Different classification methods from sklearn such as GradientBoosting, AdaBoostClassifier, Bagging Classifier, Naive Bayes(BernoulliNB), DecisionTreeClassifier, ExtraTreesClassifier, HistGradientBoostingClassifier, KNeighborsClassifier, LabelPropagation, LabelSpreading, LogisticRegression, MLPClassifiers, RadiusNeighborsClassifier, LinearSVC, RandomForestClassifier and GaussianProcessClassifier were used to predict the class of land.

Hyperparameter tuning:

The hyperparameters for some of the models were determined by Grid Search method (GridSearchCV) and they have been summarized as follows:

Model	Hyperparameters
GradientBoosting Classifier	{'n_estimators': 150, 'learning_rate': 0.1, 'max_depth': 1}
LabelPropagation	{ 'kernel': 'knn', 'n_neighbors': 7, 'max_iter': 1000 }
RandomForestClassifier	{'criterion': 'gini', 'max_depth': 4, 'n_estimators': 100}
KNeighborsClassifier	{'algorithm': 'auto', 'n_neighbors': 5, 'p': 2, 'weights': 'uniform'}

The best accuracy score was obtained with the Gaussian Process Classifier (GPC). The set of hyperparameters for our chosen model is as follows:

Hyperparameter	Value
optimizer	min_l_bfgs_b
n_restarts_optimizer	1
max_iter_predict	100
warm_start	False
multi_class	one_vs_rest

5) Implementation :

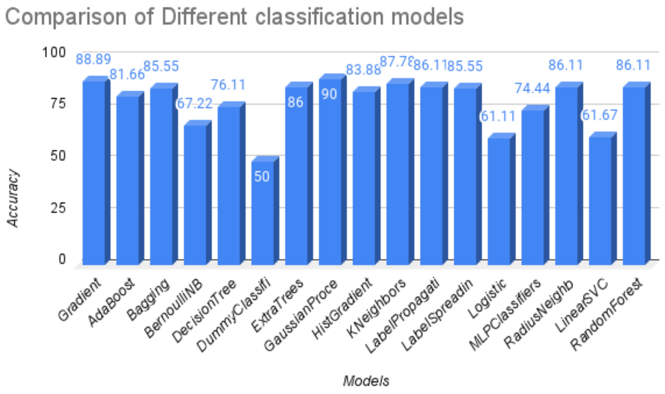
The code for the proposed work can be accessed here.

VI. RESULT ANALYSIS

Summary of the different models :

Model Name	Non-Rice f1-score	Rice f1-score	Overall Accuracy Score (%)
Gradient Boosting	0.88	0.89	88.89
AdaBoost Classifier	0.82	0.81	81.66
Bagging Classifier	0.86	0.85	85.55
BernoulliNB	0.65	0.69	67.22
DecisionTree Classifier	0.76	0.77	76.11
DummyClassifier	1.00	0.00	50.00
ExtraTrees Classifier	0.86	0.86	86.00
GaussianProcess Classifier	0.90	0.90	90.00
HistGradient Boosting Classifier	0.85	0.83	83.88
KNeighbors Classifier	0.88	0.87	87.78
LabelPropagation	0.86	0.86	86.11
LabelSpreading	0.86	0.85	85.55
Logistic Regression	0.56	0.65	61.11
MLPClassifiers	0.76	0.81	74.44
RadiusNeighbors Classifier	0.87	0.86	86.11
LinearSVC	0.56	0.66	61.67
RandomForest Classifier	0.87	0.86	86.11

As observed, Gaussian Process Classifier works best for this problem statement with an overall out sample accuracy of 90%.



V. CONCLUSION

Crop identification and crop yield forecasting are two important applications of satellite data in agriculture, which have gained significant attention in recent years due to the availability of free and open time-series satellite data, as well as advancements in machine learning and cloud computing technologies.

The use of satellite data enables the identification of crops at large spatial scales and can provide accurate information on crop type, health, and growth. This information can be used to assess crop productivity, detect potential crop stressors, and monitor changes in land use and land cover. Furthermore, machine learning algorithms can be trained on historical satellite data to develop models that can forecast crop yields based on environmental variables, such as temperature, precipitation, and soil moisture.

Such models have the potential to be a significant contribution to food security issues around the world, as they can help farmers and policymakers make informed decisions on crop management, resource allocation, and food distribution. Accurate crop yield forecasting can also help prevent food price volatility and reduce the risks associated with food insecurity.

In addition to these benefits, the availability of free and open satellite data has also democratized access to crop monitoring and forecasting tools, enabling small-scale farmers and low-resource countries to leverage these technologies to improve their farming practices and achieve greater food security.

VI. FUTURE SCOPE

The future scope of working with satellite data on crop yield classification problems in machine learning is significant, as it offers the potential to enhance crop yield prediction and management practices, leading to improved food security and agricultural sustainability.

The use of machine learning models, such as neural networks, decision trees, and support vector machines, to analyze satellite data for predicting crop yields is a promising research area.

As it is done here, these models are trained on large datasets of satellite imagery and ground truth data and can learn to recognize patterns in the data that are associated with different crop yields.

Additionally, the use of satellite data in combination with other sources of information, such as weather data, soil data, and management practices, can further improve the accuracy of crop yield predictions. This type of integrated approach can help farmers make more informed decisions about crop management and resource allocation, ultimately leading to more efficient and sustainable agricultural practices.

Furthermore, advances in remote sensing technology, including higher resolution and more frequent satellite imagery, can provide even more detailed information on crop health and yield, allowing for more precise and accurate crop yield predictions. These advancements will likely be accompanied by the development of new machine-learning algorithms that can handle these larger and more complex datasets.

In summary, the future scope of working with satellite data on crop yield classification problems in machine learning is significant and has the potential to revolutionize agriculture by providing more accurate and efficient crop yield predictions and management practices.

Acknowledgment

We would like to express our gratitude to KIIT-DU for letting us carry out this project and Dr. Chittaranjan Pradhan, Associate Professor, KIIT-DU, who has been our guide throughout.

References

[1] A Generalized Multimodal Deep Learning Model for Early Crop Yield Prediction Arshveer Kaur, Poonam Goyal, Kartik Sharma, Lakshay Sharma, Navneet Goyal; *Department of Computer Science & Information Systems*
BITS Pilani, Pilani Campus, Rajasthan, India
2022 IEEE International Conference on Big Data (Big Data)

[2] Monocot Crop Yield Prediction using Sentinal-2 Satellite Data
Laxmi B Rananavare, *Dept. of CSE, REVA University,* Bangalore, India
Sanjay Chitnis *School of CSE, RV University,* Bangalore, India
2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)
Bangalore, India.

[3] M. Ray, A. Rai, V. Ramasubramanian, and K. Singh, "Arima-wnn hybrid
model for forecasting wheat yield time-series data," *J. Ind. Soc. Agric.*
Stat, vol. 70, no. 1, pp. 63–70, 2016.

[4] L. K. Petersen, "Real-time prediction of crop yields from modis relative
vegetation health: A continent-wide analysis of Africa," *Remote Sensing*,
vol. 10, no. 11, p. 1726, 2018.

.

Authors

Ruchira Saha–Ruchira Saha, B.Tech. pre-final year, KIIT-DU, 2005259@kiit.ac.in
Devashree Baruah – Devashree Baruah, B.Tech. pre-final year, KIIT-DU, 20051066@kiit.ac.in
Sanchari Ray – Sanchari Ray, B.Tech. pre-final year, KIIT-DU, 2005264@kiit.ac.in
Oshmita Sarkar–Oshmita Sarkar, B.Tech. pre-final year, KIIT-DU, 20051583@kiit.ac.in