**LIST OF ABBREVIATIONS**

AOI - Area Of Interest

ML - Machine Learning

IR - Infrared

SWIR - Short Wave Infrared

RS - Remote Sensor

MODIS - Moderate Resolution Imaging Spectroradiometer

ESA -   European Space Agency

NDVI - Normalized Vegitation Index

AFRI - Aerosol free Vegitation Index

LAI - Leaf Area Index

SVM - Support Vector Machine

CNN - Convolutional Neural Network

DNN - Deep Neural Netwok

ANN - Artificial Neural Network

1A1 - One Against One

1AA - One Against All

# Introduction

## Background

Rice is the main food of the inhabitants of Sri Lanka. Paddy crops are cultivated as a

wetland crop in all the districts. The total land devoted to paddy is estimated to be about

708,000 Hectares at present (Division, n.d.). There are two cultivation seasons namely Maha and Yala which are synonymous with two monsoons in Sri Lanka. According to Paddy Statistics of Department of the census and Statistic, the average yield of paddy estimated for 2017 Yala season was 83.2 bushels per net acre. It decreases 2.4 bushels per net acre compare to 2016 Yala Season and Paddy estimated 2017/2018 Maha season was 83.43 bushels per net acre. It increases 123.26 bushels per acre compare to previous Maha season. The aforementioned data infers the growth of the rice yield per unit area, and also reflects the development of the national economy. Therefore, precisely mapping the cultivation is and estimating the yield of paddy rice are both crucial for national food security and national development evaluation.

In Sri Lanka, Average paddy yield is estimated by a simple survey. This process, commonly known as “crop cutting survey”, was introduced by the Food and Agriculture Organization of the United Nations (FAO) in 1951. At present, a sample of 3,000 villages for the main season (Maha) and 2,000 villages for the second season (Yala) are selected to carry out experiments. (Division, n.d.). However, this method fails to predict rice yield before the harvest as the sampling is conducted during the harvest. As well as, this survey is usually time-consuming, subjective, costly and likely to significant errors because of the limited ground survey, resulting in poor rice yield estimations.

Paddy yield forecast a few months before the harvest can be of huge importance to take necessary management decision regarding seasonal crop productions. In addition, it is most important to strategic planning and decision making regarding the food security and facilitation of safe harvest storages. According to the Sri Lankan Department of Agriculture, in the 2015/2016 Maha and 2016 Yala seasons there were 2.5 and 1.3 million metric tons of rice produced, respectively. However, during these two seasons, it was recorded that farmers in many parts of the country waited days near the storage facilities to receive payment for their crop. In the end, they were disappointed over the failure to set up an adequate mechanism to purchase their harvest. On the other hand, in early 2017, the price of rice increased significantly due to the inadequate stores of rice within the country. These two recent incidents reflect the inadequate management and poor decision making of national rice production in Sri Lanka. The conditions of surplus in 2015/2016 and scarcity in 2017 were primarily due to the unavailability of a system to forecast the expected paddy yield by a considerable time before harvesting (T. L. Dammalag).

So, to overcome previously mentioned cases by proper import, export or rearrange storage facilities, taken based on reliable yield estimations as suggested by N.A. Noureldin (N.A. Noureldin, 2013). Recently, there are lunching various satellites like Landsat, Sentinel, Tera etc with particular remote sensors to crop monitoring. Whereas, remote sensing data have the potential to provide timely, systematic high quality spatial and accurate details about land feature including the environmental impact of the crop growth (Kogan, 2002). Using sentinel-2A satellite data, creating a machine learning model to identify the stage of the paddy crop and paddy yield prediction in the ripening phase in real-time.

## Motivation Facts

* **Resourcefulness**

In the early years, there was no opensource satellite to get the NDVI values and satellite images of AOI to researches to conduct this kind of research. But nowadays we have several opensource satellites and tools to archive the bands' details and analyse them. And also there are several kinds of research were conducted in the research area. Therefore, the availability of data has been a great motivation to increase the researches in the Machine Learning field.

* **Helpfulness**

In Sri Lanka, there is poor management of rice storage because there is no proper accurate mechanism to forecast the yield of rice before the harvest. This leads to directly national economy. Because paddy rice is both crucial for national food security and national development evaluation. To overcome the improper management of rice storage(export, import), this model will help.

* **Popularity**

Rice is the single most important crop occupying 34 percent (0.77 /million ha) of the total cultivated area in Sri Lanka (RRDI, n.d.). It is the largest cultivated crop in Sri Lanka

* **Availability of Software and frameworks**

Unlike the early days at present, there are lots of software and frameworks which support machine learning. It makes easier to focus on the research, more than worrying about how to implement it. MATLAB, OpenCV, Weka are good examples of software and Scikit-learn, Tensorflow, are good examples of frameworks which support machine learning.

## Objectives

* Identify the best bands combination of Sentinel-2A satellite images to paddy crop analysis and paddy crop classification(B11, B08, B04 / B11, B08, B02 / B04, B03, B02).
* Identify the cloud images from best bands combination.
* Identify the ripening phase of AOI.
* Analysis of the relationship between the linear or poly gradient of NDVI over the paddy season vs paddy yield of that paddy season(past data).

## Benefits of the model and web-platform

* When given the coordinates of some area can be classified the paddy stage.
* Can be classified the given area contain the cloud or not.
* Can get approximate yield prediction in AOI.
* Real-time processing with satellite images

# Literature Review

## Overview of the chapter

This chapter contains the knowledge gathered to conduct the research project. This Knowledge was gathered from studying theories of machine learning, studying existing methods for image classification, regression models and the ways for optimizing them and also how spectral bands affect the vegetation and clouds such as IR, red, blue, green, SWIR and different VI(vegetation indices) for creating a model to predict the harvest of paddy.

## Introduction

This research addresses this very important and urgent issue on how to predict the harvest through NDVI values in the third phase(ripening phase) of paddy rice using remote sensing images with support vector machines . The non-destructive nature and capabilities of large scale observation and on-site monitoring makes remote sensing technique being an ideal means used for implementation practices in rice production (Chwen-Min Yan, March 2007). (Sidik Mulyono M. I., 2012) proposed a rice yield prediction method based on land and airborne hyperspectral images. It involved laborious and high-cost campaigns both infield and airborne. It was aimed as a capacity building measure to anticipate the launching of the next generation of Hyperspectral spaceborne satellite (Hyperspectral Image Suite: HISUI). In the absence of hyperspectral satellite imagery, it is very difficult to extrapolate the application of our model and thus to extend our area of estimations. Whereas, the airborne campaign is quite expensive. Instead of waiting and relying on the availability of hyperspectral satellite images, we study the use of Sentinel-2A which is a kind of multispectral remote sensing satellite operated and developed by the ESA(European Space Agency). (Ferencz C., 2004) deduce with their research that Satellite Remote Sensing (RS) can be successfully applied for this particular research area. The most multispectral satellite system measure varies spectral bands including the visible to the mid-infrared region of the electromagnetic spectrum (J.K Shwetank, 2010). The main advantage of sentinel-2A is its direct broadcast capability, and anyone can constantly download this broadcast data free of charge. Moreover, its acquiring system almost covers all of the earth surfaces, and acquire every 4-5 days at the same place, so make it possible to apply for monitoring an object on the earth surface. It plays a vital role in broad applications such as the development of the prediction model for crop yields and classification of crops using multidimensional regression of NDVI and surface temperature (Paul C. Doraiswamy, 2007).

The analysis of Sentinel-2A data for paddy growth stages classification is not a trivial task due to many factors such as large spatial variability of each growth stage due to non-uniform plantation time, atmospheric effects and the curse of dimension in dealing with a larger number of frequency bands. (Tianxiang Zhang, 2017) develops a novel approach to analyze satellite remote sensing images, particularly Sentinel-2A satellite images using machine learning techniques. (Sidik Mulyono M. I.) address the problem of harvest time and area prediction through growth stages classification using MODIS remote sensing data with support vector machines.

Regarding with yield prediction model literature, (Rajitha.K) examines the inﬂuences of cirrus cloud on spectral indices of vegetation like NDVI and AFRI. The spectral absorption typically occurs from 670 to 780 nm wavelength range of the electromagnetic spectrum (P. Kempeneers, 2004). In this case, leaf chlorophyll has a strong absorption at 0.45 μm and 0.67 μm, and high reflectance at near-infrared (0.7–0.9 μm). Healthy plants have a high NDVI (normalized difference vegetation index) value because of their high reflectance of infrared light, and relatively low reflectance of the red spectrum (J Moore, 2003). The model is based on vegetation indices (VI) which could be collected from remote sensing satellite data. VI are optical measures of vegetation canopy ‘‘greenness’’. Some researches have observed that there is a linear relationship between rice yield and VI calculated based on satellite (Cheng Q., 2010) also shows the exponential relationship (WayanNuarsa I., 2011). They give a direct measure of photosynthetic potential resulting from the composite property of total leaf chlorophyll, leaf area, canopy cover, and structure. In this case, the NDVI value is closely related to crop yield. It has a direct correlation with LAI, biomass and vegetation cover (Xiao), (Nemani, 1989). Most researches have observed that correlation between NDVI and green biomass yield, therefore NDVI can be used to estimate yield before harvesting (M.S., 1997). A research was conducted in Indonesia using Landsat, found that the best age of a paddy yield prediction is 63 days after transplanting (WayanNuarsa I., 2011). Another research in Egypt concludes that the best age of paddy yield prediction is 90 days after transplanting (N.A. Noureldin, 2013). Another research also in Sri Lanka show the best age of yield prediction is 80 days after transplanting (T. L. Dammalag).

## Image classification methods

There are several ways for image classification. Most of the classifiers are a maximum likelihood, minimum distance, decision tree, support vector machine and neural network. They are making a conclusive decision about the land cover class and require a training sample. In contrast, the clustering-based algorithm, such as K-NN, K-mean or ISODATA, are an unsupervised classifier, and fuzzy-set classifier is soft classification providing more information and potentially a more accurate result.

For identifying the stage of paddy crop, in this research, we use an image classification algorithm. Here represent some of the related work regarding RS image classification. Mainly for this research, we literature reviewed on SVM(Support Vector Machine) approach for multispectral image classification on multi-temporal data.

### SVM (Support Vector Machine)

SVMs are a recently new supervised classification technique for the land covering and mapping community. They have their fundamental in Statistical Learning Theory and have acquired prominence because they are robust, accurate and are effective even when using a small training dataset. Over the last two decades or so, RS has increasingly become a prime source of land cover information. As the name suggests, SVMs are essentially binary classifiers, however, they can be improved to handle the multiple classification tasks in remote sensing studies.

Support Vector Machines have their roots in Statistical Learning Theory (N, 1995). They have been broadly applied to machine vision fields to recognition things such as character, handwriting digit and text. (Joachims, 1998; N, 1995), and recently satellite image classification problems (Huang C., 2002). SVMs, like ANN (Artificial Neural Networks) and other nonparametric classifiers, have a reputation for being vigorous (Foody M. G., A Relative Evaluation of Multiclass Image Classification by Support Vector Machines, 2004a; Foody M. G., Toward Intelligent Training of Supervised Image Classifications: Directing Training Data Acquisition for SVM Classification, 2004b). SVMs function by nonlinearly projecting the training dataset in the input space to a feature space of higher dimension by use of a kernel function. This results in a linearly separable dataset which can be separated by a linear classifier. This process enables us to the classification of RS image datasets which are usually nonlinearly separable in the input space. In many cases, classification in high dimension feature spaces results can be over-fitting in the input space. By the way, in SVMs over-fitting is controlled by the principle called “structural risk minimization” (N, 1995). The heuristic risk of misclassification is reduced by maximizing the margin between the data points and the decision boundary (D, 2004 November). In practice, the tradeoff between these factors is managed through a margin of error parameter(usually called C parameter) that is tuned through cross-validation methods (D, 2004 November). The functions used to project the dataset from input space to feature space are sometimes called kernels machines (or kernel), examples of which include polynomial, quadratic functions and Gaussian (more commonly referred to as radial basis functions). Each function has unique parameters which have to be determined according to classification and also they are usually determined through a cross-validation process. A deeper mathematical thesis of SVMs can be found in Vapnik (1995), Campbell (2000) and Christianini (2002). SVM is basically designed for binary classification (Melgani F., 2004). But there exist strategies using that can be adopted to multiclass tasks associated with remote sensing studies.

### SVM Multiclass Strategies

SVM classification is mainly a binary classification technique, which needs to be modified to handle the multiclass tasks in real-world problems such as getting the land cover information from satellite images. There are two of the common approaches to enable this adaptation which are the 1A1 and 1AA techniques. The 1AA technique represents the earliest and rapid SVM multiclass approach (Melgani F., 2004) and involves the division of an N class dataset into N two-class cases. Means that this will train one classifier per class in total N classifiers. For class ith,  it will assume ith labels as positive and the rest as negative. Say that the classes of interest in a satellite image include vegetation, water and built-up areas, the classification would be effected by classifying vegetation against non-vegetation areas i.e. (water and built-up areas) or water against non-water areas i.e. (vegetation and built-up areas). The 1A1 technique involves constructing a machine for each pair of classes resulting in N(N-1)/2 machines. Means that 1A1 have to train a separate classifier for each different pair of labels. This leads to N(N-1)/2 classifiers. From ML theory, it is known that the disadvantage the 1AA approach has over 1A1 is that its performance can be compromised due to unbalanced training datasets (Gualtieri J. A., 1998). However, 1A1 is much less sensitive to the problems of unbalanced datasets but is much more computationally expensive. Some researches concluded that whereas one can be certain of high classification results with the 1A1 approach, the 1AA yields approximately as good classification accuracies (Gidudu Anthony).

**REFERENCES**

Cheng Q., W. X. (2010). Mapping paddy rice yield in Zhejiang Province using MODIS spectral index. .

Chwen-Min Yan, R.-K. C. (March 2007). Differences in Growth Estimation and Yied Prediction of Rice Crop Using Satellite Data Simulated From Near Ground Hyperspectral Reflectance. *Journal of Photogrammetry and Remote Sensing*.

community, w. (n.d.). *Chlorophyll*. Retrieved from wikipedia: https://en.wikipedia.org/wiki/Chlorophyll

community, w. (n.d.). *Satellite\_crop\_monitoring*. (wikipedia) Retrieved from https://en.wikipedia.org/

D, M. (2004 November). Comparing SVM and GMM on parametric feature-sets. *In Proceedings of the 15th Annual Symposium of the Pattern Recognition Association.* Cape Town, South Africa.

Deng N., e. a. (2009). Support vector Machine Theory, algorithms and Development. *Science Press, Beijing*, 176.

Division, A. a. (n.d.). *paddy statistics*. (Department of Census and Statistics, Colombo, Sri Lanka.) Retrieved from http://www.statistics.gov.lk/agriculture/Paddy%20Statistics/PaddyStats.htm

Ferencz C., B. P.-Á. (2004). Crop yield estimation by satellite remote sensing . *International Journal of Remote Sensing*, 25, 4113-4149.

Foody M. G., a. M. (2004a). A Relative Evaluation of Multiclass Image Classification by Support Vector Machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42, 1335-1343.

Foody M. G., a. M. (2004b). Toward Intelligent Training of Supervised Image Classifications: Directing Training Data Acquisition for SVM Classification. *Remote Sensing of Environment*, 93, 107-117.

Gidudu Anthony, H. G. (n.d.). *Classification of Images Using Support Vector Machines.* Department of Electrical and Information Engineering, University of the Witwatersrand.

Gualtieri J. A., a. C. (1998). Support vector machines for hyperspectral remote sensing classification. *In Proceedings of the 27th AIPR Workshop: Advances in Computer Assisted Recognition.* Washington.

Huang C., D. L. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23, 725–749.

J Moore, N. H. (2003). Examining the development of a potato crop nutrient management trial using reflectance sensing. *ASAE Annual International Meeting. Las Vegas, Nevada, USA.* Las Vegas, Nevada, USA.

J.K Shwetank, K. B. (2010). Review of rice crop identification and classification using hyper spectral image processing system. *International Journal of Computer Science & Communication 1*.

Joachims, T. (1998). Text categorization with support vector machines—learning with many relevant features. *In Proceedings of the 10th European Conference on Machine Learning*, (pp. 137–142). Chemnitz, Germany.

Kogan, W. T. (2002). Monitoring Brazilian soybean production using NOAA/AVHRR based vegetation condition indices. *International Journal of Remote Sensing*.

M.S., R. (1997). Operational yield forecast using AVHRR NDVI data: reduction of environmental and inter-annual variability. *International Journal of Remote Sensing*, 18 (5), 1059–1077.

Melgani F., a. B. (2004). Classification of hyperspectral remote sensing images with Support Vector Machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42,1778 – 1790.

N, V. V. (1995). *The nature of statistical learning theory.* New York.

N., X. (2011). Comparison of multi-class support vector machines. *Computer Engineering and*.

N.A. Noureldin, M. A. (2013). Rice yield forecasting models using satellite imagery in Egypt. . *The Egyptian Journal of Remote Sensing and Space Science*.

Nemani, R. R. (1989). Testing a theoretical climate-soil leaf area hydrological equilibrium of forests using satellite data and ecosystem simulation. *Agriculture and Forest Meteorology*, 44, 245–260.

P. Kempeneers, D. B. (2004). Wavelet based feature extraction for hyper spectral vegetation monitoring,. *Image and Signal Processing for Remote Sensing* .

Paul C. Doraiswamy, B. A. (2007). Operational Prediction of Crop Yields Using MODIS Data And Products, International Archives of Photogrammetry Remote Sensing and Spatial Information Scince Special Publication.

Rajitha.K, P. M. (n.d.). Effect of Cirrus Cloud on Normalized Difference Vegetation Index (NDVI) and Aerosol Free Vegetation Index (AFRI): A Study Based on LANDSAT 8 Images.

RRDI. (n.d.). *rrdi*. (Department of Agriculture) Retrieved from https://www.doa.gov.lk/rrdi/index.php?option=com\_sppagebuilder&view=page&id=42&lang=en

Sidik Mulyono, M. I. (2012). Genetic Algorithm Based New Sequence Principal Component Regression (NS-PCR) For Feature Selection And Yield Prediction Using Hyperspectral Remote Sensing Data. *International Geoscience and Remote Sensing Symposium, 2012* .

Sidik Mulyono, M. I. (n.d.). A Paddy Growth Stages Classification Using MODIS Remote Sensing Images with Balanced Branches Support Vector Machines . *ICACSIS 2012*.

T. L. Dammalag, P. M. (n.d.). MODIS Satellite Data Based Rice Yield Forecasting Model for Sri Lanka: A Pilot Study in Kurunegala District. *Asian Journal of Geoinformatics*.

Tianxiang Zhang, J. S.-H. (2017). Band Selection in Sentinel-2 Satellite for Agriculture Applications. *Proceedings of the 23rd International Conference on Automation & Computing, University of Huddersfield,Huddersfield, UK, 7-8 September 2017.* China.

WayanNuarsa I., N. F. (2011). Rice Yield Estimation Using Landsat ETM Data and Field Observation. *Journal of Agricultural Science 4* .

Xiao, X. B. (n.d.). Mapping paddy rice agriculture in southern China using multi-temporal MODIS images. . *RemoteSensing of Environment* , 95, 480–492.

Zhang Y., Z. Y. (2011). Application of Least Squares Support Vector Machine in Fault Diagnosis. *ICICA*, (pp. 192–200.). Springer, Heidelberg .

Zhang, Y. (2012). *Support Vector Machine Classification Algorithm.* Tangshan, China: Qinggong College, Hebei United University.