ABSTRACT

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

MSI - multi-spectral instrument

AOD - Aerosol Optical Depth

DOS - Dark Object Subtraction

# 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 compared to the 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, the 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 make necessary management decisions 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 crops. 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 analyze 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 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 a 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 that 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 the 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 identified in the paddy stage.
* It can identify the given area contain the cloud or not.
* Can get approximate yield prediction in AOI.
* Real-time processing with satellite images

## Overview of the Research

# 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 various 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's 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).

## Related work

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 yield prediction, (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 the 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 shows the best age of yield prediction is 80 days after transplanting (T. L. Dammalag). Some studies identify that the Potential Bands of Sentinel-2A Satellite for Classification Problems in Precision Agriculture is for index-based classification (NDVI, NDWI), index-related band based classification (Band 4, Band 8A, Band 11), MI scored band based classification (Band 6, Band 7, Band 8, Band 8A, Band 9, Band 5 and Band 3) have more details (Tian-Xiang Zhang, 2018).

## 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, is an unsupervised classifier, and the 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 for RS Image Classification

#### Overview

SVMs are a recently new supervised classification technique for the land covering and mapping community. They have their fundamentals 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 a 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 that 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 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. This 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. This means that 1A1 has 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 good classification accuracies (Gidudu Anthony).

### Grid Search Method

A grid search approach is an alternative for finding the best model parameter to predict unlabelled data accurately (M. Ataei, 2004). The method is known as a systematic method for the best parameter values, each of which should be tested by setting some prediction values. The Method then displays the score to be chosen for each parameter value. This approach is applicable when the amount needed for each of the separate variables is known to be within limits specified by upper or lower limits (M. Ataei, 2004). The Hsu, Chang and Lin (C.W. Hsu, 2003) search method shows that the best value in two parameters, which are, in this case, c and y, is the grid search method and the cross-validation. The research investigates different pairs of (C, y) and is selected for the most detailed cross-validation.

### Data Augmentation

The goal of Data Augmentation is to extend the dataset to resolve data representation limitations and reduce the problem of overfitting. It increases the efficiency of the model and avoids unbalanced learning. Data increases were commonly reported in many areas to mitigate data scarcity (Ding, 2016). The technique also eliminates over-fitting, which can occur when adding model-related data instances and over-complex models (Wang J. P., 2017). Sadly, there was not much work recorded with data increase techniques in the satellite image domain.

The most widely used image transformations are in satellite image clipping, spinning, flipping, moving, sorting and translating (Ding, 2016; Wang J. P., 2017), while materials were not available in the sense of satellite image super-resolution to form a strong opinion on which augmentation techniques were more useful.

### Regression analysis

The statistical technique used for related variables (Bowerman, 1990) is the analysis of regression. The central purpose is to construct a mathematical model that connects dependent variables with separate variables. In general, a single algebraic equation of the form (Draper, 1981) is used to defy a regression model.

Three forms of regression models are available. This consists of

1. the grade-day variant of the equation (VBDD)
2. the linear regression model
3. the change-point models.

All of them have extensive regression of lower squares in order to identify the model coefficients (Kissock etc., 2003).

### Atmospheric Correction

When reflecting the surface of the ground the correction of the Atmosphere must be done to correct atmospheric effects in satellite scenes. The atmosphere influences the radiation signals recorded by satellite sensors through the diffusion of light, absorbance and refractory. The accurate recuperation of land-surface reflectiveness becomes more important with the development of quantitative remote sensing, and an acceptable accuracy atmosphere correction algorithm and easy application are urgently needed.

There have been numerous efforts to develop atmospheric image correction algorithms. Radiographical communication codes (for example, 6S and Modtran) (ZAGOLSKI, 1995; LIANG, 2001) are often used for more advanced methods. The research showed that the digital image number could precisely be converted to a land surface reflectance by these radio transfer codes (HOLM, 1989). Nevertheless, these corrections require in situ data in the image scene or simultaneous satellite passage measurements of air-optic properties. For many remote sensing applications this is inappropriate and always unlikely, particularly where historical data are used. In order to overcome these restrictions, a range of other approaches have also been suggested, for instance the dark object subtraction method (DOS) (MORAN, 1992; CHAVEZ, 1989). Invalid object method (HALL, 1991) Histogram matching method (RICHTER, 1996). The DOS approach is commonly used, purely photos based, with no in situ calculation evidence (SOUDANI, 2006). Nonetheless, the two important factors used to measure transmissions in the DOS model are the aerosol optical depth (AOD) and the overall ambient water vapor output.

# Materials and methods

## Overview of the chapter

This chapter includes the process of systematically implementing the concept in a scientific way. This chapter also includes the ideas, equations, methods and systems that help to carry out this research.

## Tools and Frameworks

### Python(v3.7.5)

According to the official Python documentation, Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum from 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). Python is interpreted, interactive and object-oriented language. The features of Python are:

* Easy to learn.
* Easy to read.
* Easy to maintain.
* It consists of a broad and standard library.
* Python has support for an interactive mode, which allows interactive testing and debugging of snippets of code.
* Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
* Python is an extendable language.
* Python provides interfaces to all major commercial databases.
* Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
* Python provides a better structure and support for large programs than shell scripting.
* It supports functional and structured programming methods as well as OOP.
* It can be used as a scripting language or can be compiled to byte-code for building large applications.
* It provides very high-level dynamic data types and supports dynamic type checking.
* It supports automatic garbage collection.
* It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java

(foundation, 2019)

### NumPy(v1.17.3)

NumPy is a general-purpose array-processing package designed to efficiently manipulate large multi-dimensional arrays of arbitrary records without sacrificing too much speed for small multi-dimensional arrays. NumPy is built on the Numeric code base and adds features introduced by num array as well as an extended C-API and the ability to create arrays of arbitrary type which also makes NumPy suitable for interfacing with general-purpose database applications. There are also basic facilities for discrete Fourier transform, basic linear algebra and random number generation. All Numpy wheels distributed from PyPI are BSD licensed. The used NumPy version is v1.17.3 for this research. (Developers, 2019)

### Pandas(v0.25.3)

Pandas is an open-source, BSD-licensed library providing high-performance, easy-to-use

data structures and data analysis tools for the Python programming language. pandas is a

NumFOCUS sponsored project. This will help ensure the success of the development of pandas as a world-class open-source project and makes it possible to donate to the project. Used Pandas version 0.25.3 for this research. (source, 2019)

### Matplotlib(v3.1.1)

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hard copy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code. For example, see the sample plots and thumbnail gallery. For simple plotting, the Pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc., via an object-oriented interface or via a set of functions familiar to MATLAB users. For this research, we used Matplotlib version 3.1.1. (John Hunter, 2012)

### Seaborn(v0.9.0)

Seaborn is a library for making statistical graphics in Python. It is built on top of Matplotlib and closely integrated with pandas data structures. Here is some of the functionality that seaborn offers:

* A dataset-oriented API for examining relationships between multiple variables.
* Specialized support for using categorical variables to show observations or aggregate statistics.
* Options for visualizing univariate or bivariate distributions and for comparing them between subsets of data.
* Convenient views onto the overall structure of complex datasets
* High-level abstractions for structuring multi-plot grids that let you easily build complex visualizations.
* Concise control over Matplotlib figure styling with several built-in themes
* Tools for choosing color palettes that faithfully reveal patterns in your data

Seaborn aims to make visualization a central part of exploring and understanding data. Its dataset-oriented plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. (Waskom, 2018)

### Scikit-learn(v0.21.3)

Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. Scikit-learn was initially developed by David Cournapeau as a Google Summer of Code project in 2007. Later Matthieu Brucher joined the project and started to use it as apart of his thesis work. In 2010 INRIA got involved and the first public release (v0.1 beta) was published in late January 2010. The project now has more than 30 active contributors and has had paid sponsorship from INRIA, Google, Tinyclues and the Python Software Foundation.

The functionality that Scikit-learn provides include:

* Regression
* Classification
* Clustering
* Model selection
* Preprocessing

(development, n.d.)

### Anaconda Cloud

According to Anaconda documentation, Anaconda Cloud is a package management service that makes it easy to find, access, store and share public and private notebooks, environments, and Conda and PyPI packages. Cloud also makes it easy to stay current with updates made to the packages and environments that are being used. (Anaconda, 2019)

### Conda

Official Anaconda documentation mentions that Conda is an open-source package management system and environment management system for installing multiple versions of software packages and their dependencies and switching easily between them. It works on Linux, macOS and Windows, and was created for Python programs but can package and distribute any software.

Conda is included in Anaconda and Miniconda. Conda is also included in the Anaconda subscriptions of Anaconda, which provide on-site enterprise package and environment management for Python, R, Node.js, Java, and other application stacks. Conda is also available on PyPI, although that approach may not be as up-to-date (Anaconda, 2019).

### Anaconda Navigator

According to official Anaconda documentation, it is a desktop graphical user interface included in Anaconda that allows you to launch applications and easily manage Conda packages, environments and channels without the need to use command line commands (Anaconda, 2019). Figure 1 illustrates the Anaconda Navigation Dashboard.

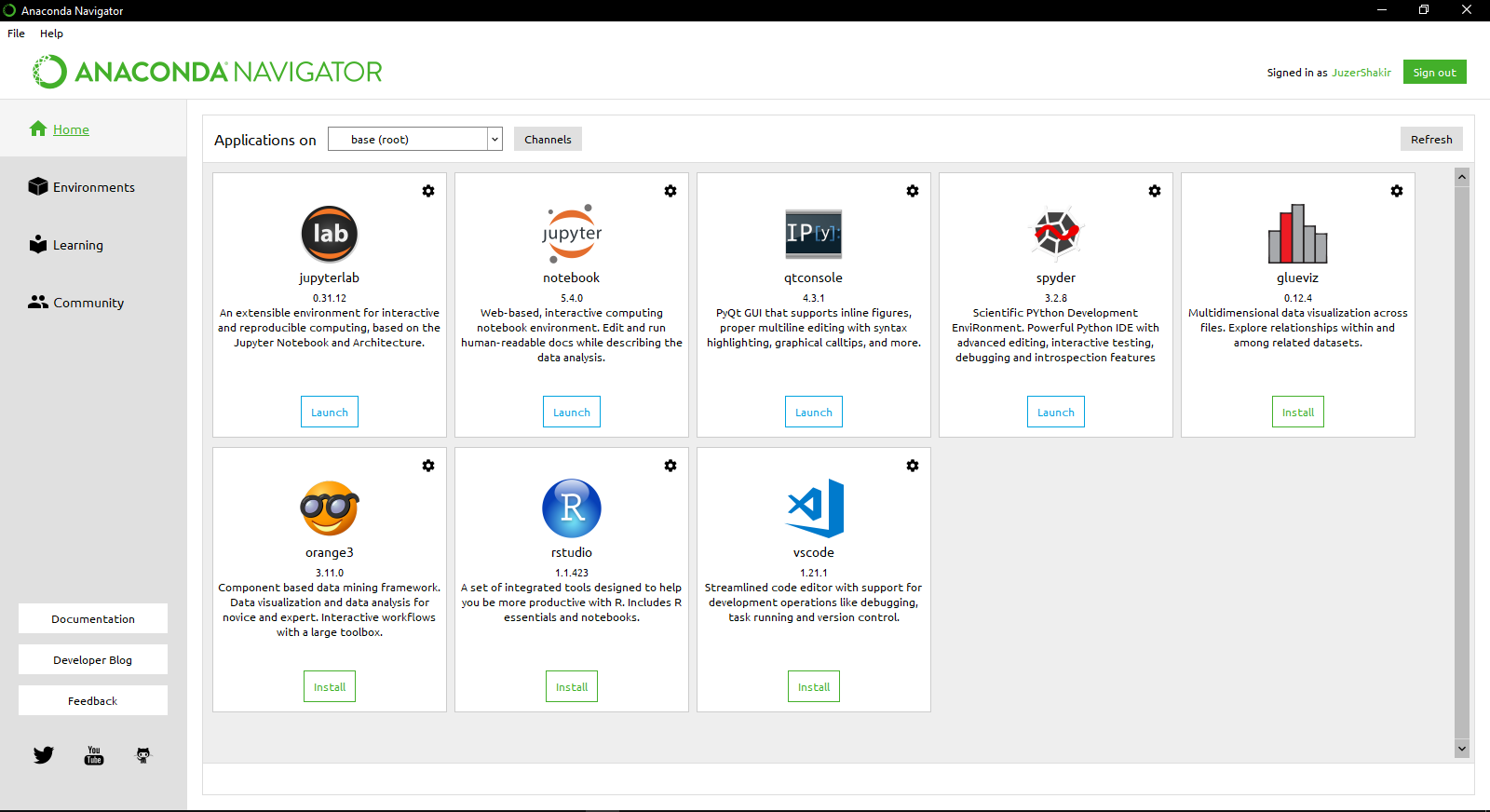


Figure : Screenshot of anaconda navigator

### Jupyter Notebook(v1.0.0)

The Jupiter Notebooks’ official documentation says that The Jupyter Notebook is an open-source web application that allows anyone to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more (Jupyter, 2019)

### Sentinel Hub(v2.6.0)

According to the Sentinel Hub web site, Sentinel Hub is a multi-spectral and multi-temporal big data satellite imaging service capable of fully automated archiving, real-time processing and distribution of remote sensing data and related EO products. Users can use APIs to access satellite data from full repositories in a matter of seconds over their AOI and limited time range.

The Sentinel Hub API is a RESTful API interface to various satellite imagery archives. It provides access to raw satellite data, rendered images, statistical analysis and much more. (Sinergise, 2019). Figure 2 illustrates the Sentinel Hub API Dashboard.

A screenshot of a computer

Description automatically generated

Figure : sentinel hub configuration dashboard

### ASUS VivoBook Pro 15 N580GD

According to the users’ Manual, the important hardware configuration of the laptop is:

Processor: Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz (12 CPUs), ~2.2GHz

Memory: 12GB

Hard drive: 1TB, Toshiba (HDD)

Graphics: NVIDIA GeForce GTX 1050

Screen: 15.6" FHD (1920 x 1080) resolution

(Asus, 2019)

The Operating System installed is Microsoft Windows 10 (64 bit) version.

## Sentinel 2 Satellite

Sentinel-2 is a Copernicus program Earth Observation Mission that systematically acquires optical imaging over land and coastal waters in a high spatial resolution (10-60 m). It has two two twin satellites, Sentinel-2A and Sentinel-2B. (community w. , Sentinel-2, n.d.)

Sentinel-2 satellite used to archive the RS data for this study. Which has [Multi-spectral](https://en.wikipedia.org/wiki/Multispectral_image) data with 13 bands in the [visible](https://en.wikipedia.org/wiki/Visible_spectrum), [near-infrared](https://en.wikipedia.org/wiki/Infrared#Regions_within_the_infrared), and [short wave infrared](https://en.wikipedia.org/wiki/Infrared#Regions_within_the_infrared) part of the [spectrum](https://en.wikipedia.org/wiki/Electromagnetic_spectrum). Revisiting every 5 days under the same viewing angles. At high latitudes, Sentinel-2 swath overlap and some regions will be observed twice or more every 5 days, but with different viewing angles (community w. , Sentinel-2, n.d.). For this study, used five surfaces, reflectance bands. They are 10m spatial resolution surface reflectance band2(490 nm), band3(560 nm), band4(665 nm) and band8(842 nm) and also 20m spatial resolution surface reflectance band11(1610 nm).

### Sentinel 2A

The mission of the Copernicus Sentinel-2A consists of a constellation of two polar-orbiting satellites placed in the same sun-synchronous orbit at 180 °. This seeks to track variation of terrestrial surface conditions and its wide swath range (290 km) and high revision time (10 days at the equator at one satellite and 5 days at 2 satellites under cloud-free conditions resulting in 2-3 days at mid-latitudes) would enable tracking of Earth's surface changes. The geographical limits are from 56 ° south to 84 ° north latitudes. The launch of Sentinel-2A, the first satellite, took place on a Vega launch vehicle on 23 June 2015 at 01:52 UTC. (ESA, 2019)

SENTINEL-2A satellite weighs approximately 1.2 tonnes. VEGA European launchers were launched. The duration of the satellite is 7.25, including a maximum of 3 months in space. 12 years of operations, including the completion of life-threatening exercises, have been supplied with batteries and propellants. (ESA, 2019)

The MSI is working passively by extracting sunlight from the earth. Once the spacecraft travels along the orbital path, new data is collected on the instruments. A filter separates the incoming light beam and concentrates on two different flags within the unit, one for the Visible and Near-Infra-Red (VNIR) ranges and another for the Short Wave Infra-Red (SWIR) bands(Figure 3). The spectral isolation of each ribbon is carried out by strip filters on top of the detectors into different wavelengths (ESA, 2019). Figure 4 illustrates the schematic view of the deployed SENTINEL-2A Spacecraft.

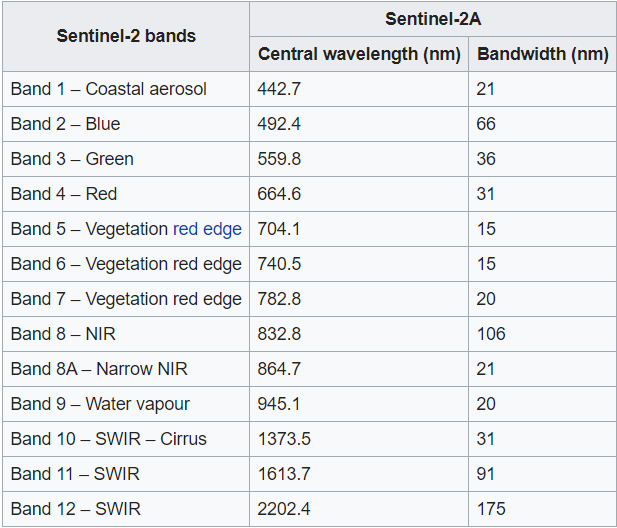


Figure :Spectral bands for the Sentinel-2 sensors

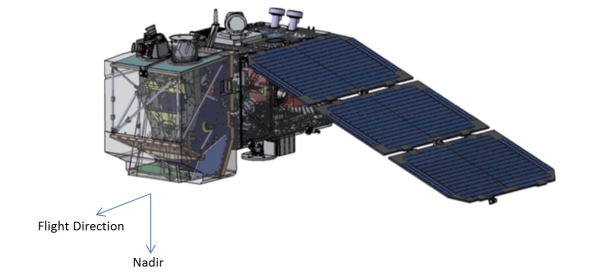


Figure : Schematic View of the Deployed SENTINEL-2 Spacecraft (image credit: EADS Astrium)

## Methodology

In this study, the main object is forecasting the paddy yield in AOI within the third phase(ripening phase) of the paddy stage based on Sentinel-2 RS data. Therefore mainly, there are three main aims that we need to do.

* Identify the best sentinel 2 bands combination for the RGB channels to paddy stage classification.
* Identify third paddy crop stages(ripening phase) using satellite images.
* Build the paddy yield forecasting model using NDVI seasonal values.

### Data Gathering

Data collecting consist of mainly three parts

* Decomposed RS images acquiring through Sentinel 2A satellite of AOI
* Corresponding temporal NDVI data points acquired through Sentinel 2A satellite
* Local paddy data gathering of AOI

#### Area of Interest(AOI)

The paddy area selected for the present study for collecting data is located in Kiridiwita, Gampaha Sri Lanka(Figure 5). This belongs to Ketawala Project. Ketawala project is about 980 acres and has thirteen farmers organizations. The initial experiment performs in Kiridiwita farmers' organization. The selected area WGS84 coordinates are 79.96983110904694, 7.084400873410644, 79.97372031211853, 7.08574770774949. This paddy area is about 15.92 acres and uniformly transplanting the rice crops in this area. The climatic condition of this study area has major changes in weather occurring during the monsoons from May to August and October to January every year. With the expected heavy rains during these two periods of the year, two seasonal rice crops are always cultivating in this area.

Reasons for selecting this particular AOI:

* Easy to reach.
* Easy to study.
* Farmers are uniformly transplanting in the paddy area.
* Same rice plants were planting().
* Data availability.

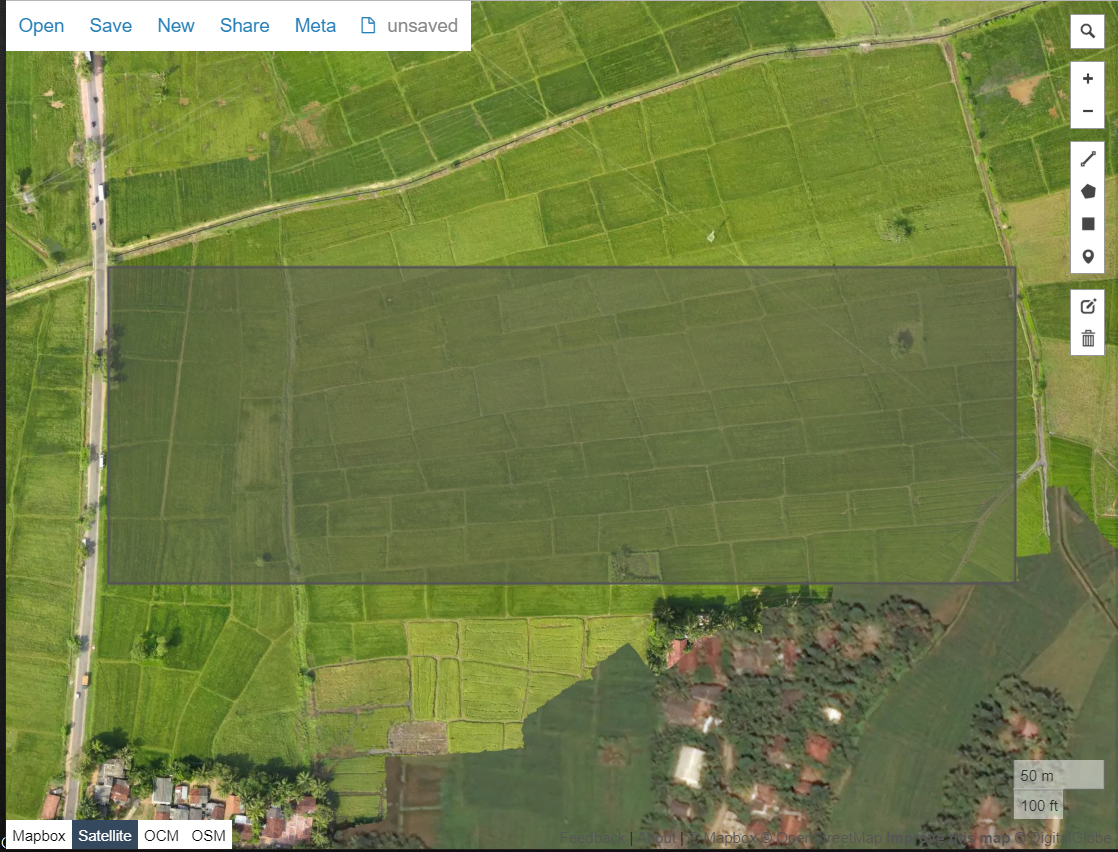


Figure : AOI

#### Gathering Satellite Data

Sentinel Hub API used to gather the decomposed RS images and NDVI values in AOI. For building a paddy classification model, surface reflectance band combinations

* (band4, band3, band2)
* (band11, band8, band4)
* (band11, band8, band2)

were selected. Here band2 is blue, band3 is green, band4 is red, band8 is IR and band11 is SWIR(Figure 8). (band4, band3, band2) typically known as “True-colour” (community w. , Sentinel-2, n.d.). which is a visible RBG color. Healthy vegetation reflects NIR(near-infrared) and green bands compare to other wavelengths. And more absorb red and blue bands (H Croft, 2017). SWIR(short wave infrared) used for crop identification (Sruthi Swathandran, 2019). Therefore, (band11, band8, band4) (Tian-Xiang Zhang, 2018) and (band11, band8, band2) were selected rather than used only visible bands. Each band's combination is applied to RGB channels and made the images. There are one hundred and seventy-eight images from 2015 to 2019 that were acquired of AOI of each band combination through the sentinel-hub cloud (Sinergise, 2019) also Corresponding temporal NDVI data points of AOI were gathered using Sentinel Hub API using below eval script code.

|  |  |  |
| --- | --- | --- |
|  |  |  |

For all querying and gathering decomposed RS images and NDVI values, used DOS1 Atmospheric correction

##### DOS1(Dark Object Subtraction 1)

DOS1 is the simplest DOS model. This states that atmospheric transmission in the direction of illumination and in the direction of viewing () is unit (no loss of atmospheric transmission) and that downwelling diffuse irradiance () is negligible (CHAVEZ, 1989). DOS1 has the form:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

DOS1 adjusts the path radiance, solar irradiance and solar zenith angle effects. This method performs well in the visible spectrum, especially for the shorter wavelength blue band, where attenuation is primarily due to dispersion. In the Near-Infrared (NIR) and Short-Wave Infrared (SWIR) region, atmospheric scattering is minimal and water vapor absorption is dominant; since DOS1 only corrects for the additive scattering effect and not for the multiplicative transmitting effect, its accuracy in the NIR and SWIR region is not acceptable.

#### Local Paddy Data Gathering

Seasonal paddy stages details were gathered by the famous organization in Kiridiwita, Gampaha and the farmers related in AOI according to the water released date of the Department of Irrigation Yakkala, Sri Lanka. Figure 6 illustrates the graphical visualization of the paddy stages’ dates changing over time from 2015 to 2019 in AOI. And the paddy past harvest details were gathered by the Department of Agriculture in Galahitiyawa. Table 1 illustrates the past harvest details in AOI Bushels per acre.

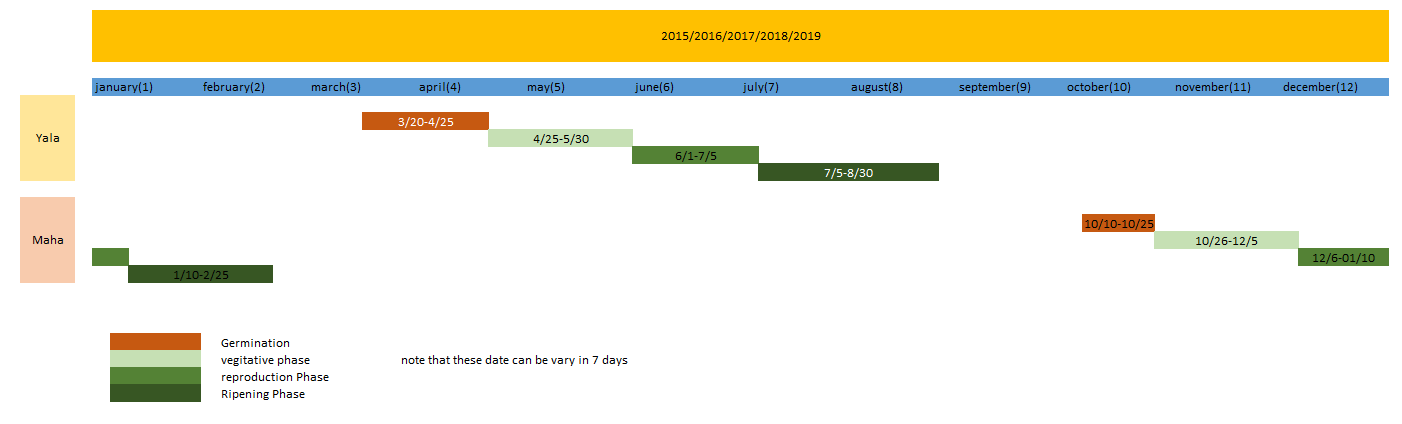


Figure : paddy stage dates 2015-2019 in AOI

|  |  |  |
| --- | --- | --- |
|  | **Yala** | **Maha** |
| 2015 | 1.28 | 1.35 |
| 2016 | 1.30 | 1.36 |
| 2017 | 1.34 | 1.36 |
| 2018 | 1.32 | 1.34 |
| 2019 | 1.32 | Still processing |

Table : paddy harvest in AOI Bushels per acre

### Data preprocessing

#### Image preprocessing

In this study, there are two main steps that need for the RS images preprocessing. They are:

* Remove the cloud covered images or white satellite images
* Labeling the images

Figure 7 and figure 8 shows the cloud image and good image respectively in bands combination (band11, band8, band2).



Figure : good image in AOI

Figure : cloud image in AOI

And figure 9 illustrates the white satellite images





Figure :white satellite sample images

The reason for the white satellite images is that we were querying the satellites that have 20% below of the cloud coverage. But for AOI never have cloud coverage below 20% on some dates (sentinel-hub-faq, 2019). By the way, we have to remove those figure 7 and figure 9 images for paddy stage classification. This was done manually for each band's combination. Once removing the unnecessary images then label the images as Germination or Vegetative phase or Reproductive phase or Ripening phase according to Figure 6 dates with corresponding satellite images’ dates. This was done for each band's combinations. By doing these main 2 steps, we were able to get 20 images per each class in each bands’ combinations.

#### Separating Training Images and Testing Images

Since this study was following a supervised learning approach, a training data set and a testing data set were needed in order to train an SVM model. Therefore, after getting rid of useless images, the total image data set was separated into two categories called training data set and testing data set. For each rice phase, 80% of images were allocated into the training image data set, and 20% to testing image data set using train\_test\_split() function in sklearn.model\_selection library (developers, 2019).

#### NDVI values preprocessing

For analysis, only get the NDVI values that have the date of good satellite images. All other NDVI values were eliminated because they have incorrect reflectance values absorb by different objects like clouds. They might give an invalid result.

### Model for paddy stage identification

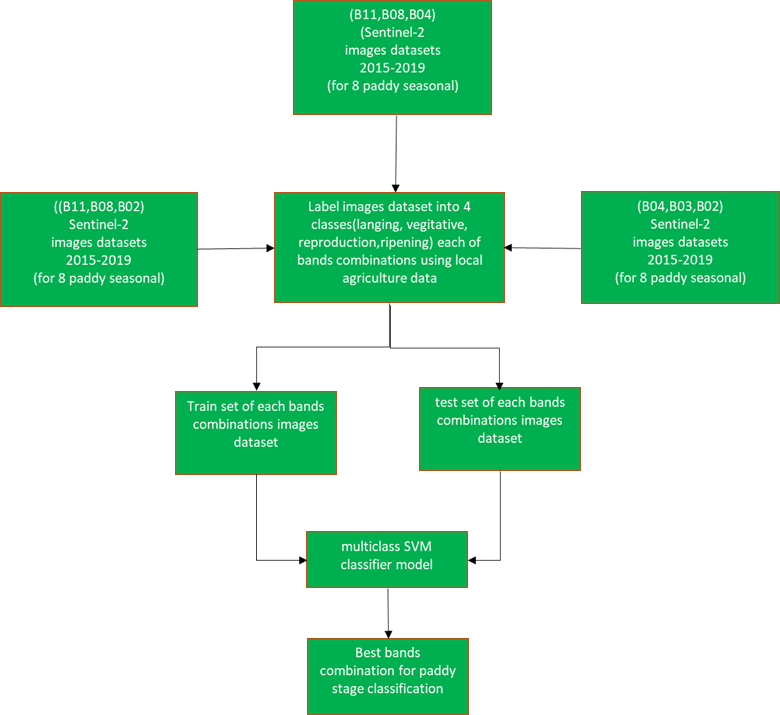


Figure : Paddy Stage Identification Diagram

The second aim is to propose a method to identify the paddy crop stages. To do that it needs to be accomplished the first aims. . For that, the image classification approach was used. Since there are very few images, the support vector machine classification model was selected (Sidik Mulyono M. I.). Figure 10 diagram illustrates the whole process of paddy stage identification.

#### Process for paddy stage identification

1. Querying decomposed RS images from Sentinel Hub API. The decomposed bands combinations are:
   1. (band4, band3, band2)
   2. (band11, band8, band4)
   3. (band11, band8, band2)

Which located in local computer harddisk separated 3 locations. The factors that were considered while querying and downloading the decomposed RS images.

* Cloud coverage below 20%
* Mosaic order is most recent
* DOS1 atmospheric correction
* 60m resolution
* Image formate is .png
* 960x330 image dimension.

1. Image preprocessing
   1. Remove the cloud covered images or white satellite images
   2. Labeling the images.
   3. Split the train images and test images
2. Train each bands’ combination with the SVM multi classifier model.
3. Find the best bands’ combination among (band4, band3, band2), (band11, band8, band4) and (band11, band8, band2) using parameter tuning.
4. Identify the paddy stage using bands’ combination of given decomposed satellite image.

#### SVM classification method

SVM is a two-dimensional description of the optimal surface evolved from the linearly separable case, the basic idea can be used in figure 1. Figure 11 shows the basic idea. Two types are separated by H without errors. H1 and H2 are places that pass the recent point of H. The distance of H1 and H2 was called a class interval. An optimal separating surface is not only to ensure error-free separation of the two types of samples but also called the largest class interval.

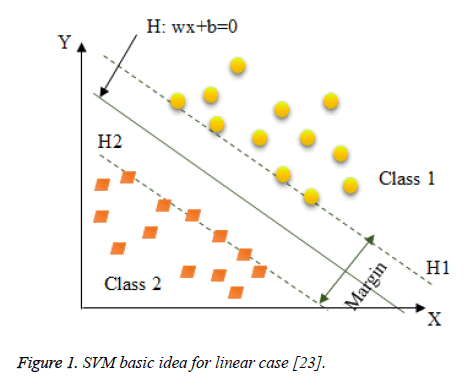


Figure : Optimal separating surface

In pattern recognition, the linear discriminate function in the n-dimensional space:

The classification hyperplane equation can be written . Linear separable case, the discriminate function was normalized. So that all training samples are met , even to meet away from the classification of the surface of the sample , so the class interval equivalent to , thus making the interval on the equivalent to or . Make a classification of the surface of all samples correctly classified, it is necessary to meet:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

Satisfy the above equation (2) and make the smallest classification surface is the optimal classification surface. A point on the hyperplane are called support vectors, they support the optimal classification surface (Deng N., 2009).

#### Parameter tuning

For hyperparameter tuning, GridSearchCV was used. There were 2 main reasons for choosing GridSearchCV for hyperparameter tuning in SVM

1. The probability that you will feel uncertain using methods to prevent a thorough search for approximation parameters.
2. Less computational time.

The following table illustrates the tuning hyperparameters for each kernel functions

|  |  |  |
| --- | --- | --- |
| **kernel** | **Regularization (C)** | **gamma** |
| linear | 1  10  100  1000 |  |
| rbf | 1  10  100  1000 | 0.001  0.0001 |

Table : parameter Grid for GridSearchCV

### Model for paddy yield prediction.

The third main aim is to build the paddy yield forecasting model using NDVI seasonal values. So that, the rice yield forecasting model was derived based on the average NDVI and maximum NDVI and cumulative NDVI in paddy seasonal. For this experiment 6 paddy seasonals were selected paddy seasons from 2015 to 2018 to build the model and 2 paddy seasonal (2019 Yala & Maha) was selected to do the accuracy assessment. Then the relationship between the average VI and yield was modeled using the recorded average rice yield data from 2015 to 2018.

I need to write the process.

“Greenness” of paddy plants is corresponding to the surface reflectance values of satellite images and it is changing according to the temporal change of paddy plants. Therefore in AOI, obtain average, maximum and cumulative of NDVI values over 90 days time period. Models were derived with a time interval of sixteen 5 days, for only 3 different stages of paddy growth.

A comparison of the results of the accuracy assessment achieved by linear and exponential models defined the best model for the relationship between rice yields and vegetation indices. The remaining average rice yield data 2019 Yala and Maha, not used for modeling, shall be used for the accuracy validation of the derived models. As a result of taking cumulative NDVI based on the average NDVI, a further rice return prediction model was drawn in addition to the NDVI model. Both the original NDVI time profile and the smooth NDVI profiles obtained via moving average technique have been used to calculate cumulative NDVI values.

#### Vegetation Indices (VI)

Vegetation Indices (VI) is used to minimize the result of the vegetation canopy radiometric response by factors such as sun-illumination, optical properties of the soil background, view geometries and meteorological factors that were discussed by Leblon (1997). Vegetation Indices (VI) are measures of vegetation “greenness” acquired by combining the results of surface reflectance measurements of the vegetation canopy in different spectral bands (Myneni R. B., 1995). The vegetation indices work based on the principle that vegetation reflects and absorbs different amounts of electromagnetic radiation in electromagnetic spectral. Vegetation indices are not characteristic of one particular physical property of the vegetation, but they constitute a combination of properties such as canopy cover, leaf chlorophyll content and leaf area (Cheng Q., 2010). Many researchers developed a number of different vegetation indexes to determine the greenness of the plant.

RVI(Ratio Vegetation Index) is the first ratio base vegetation index (Pearson R.L., 1972). It is the ratio between near-infrared and red bands. (equation 3)

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

(Rouse, 1974) improved the vegetation index and derived Normal Difference Vegetation Index(NDVI) based on near-infrared and red spectral bands as shown in equation 4

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

Green healthy vegetation more absorbs red band and weak absorbs the near-infrared band (H Croft, 2017).