# FORECASTING AGRICULTURAL CROP YIELD VARIATIONS USING BIG DATA AND SUPERVISED MACHINE LEARNING

Kariyawasan Jalath Thanthrige Dimuthu Lakmal

(188004E)

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

March 2020

## DECLRATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters dissertation under my supervision.

Signature of the supervisor: Date:

# ABSTRACT

The government of Sri Lanka is struggling to make appropriate policy decisions regarding paddy cultivation due to absence of accurate and timely data to estimate the paddy yield, land usage for paddy cultivation and area affected by various paddy diseases. Remote sensing data based machine learning implementations can be identified as a potential solution for the above issue, as remote sensing data can be used for accurate and timely estimations. However, the traditional remote sensing data resources have failed to generate accurate estimates regarding cultivated paddy extent estimations. In this study, novel optical remote sensing data resources and a hybrid approach are employed to mitigate previously reported issues. Furthermore, a multi-temporal approach is used instead of traditional mono-temporal approach by leveraging deep neural networks. This study also consists of a comprehensive comparison on novel optical remote sensing data resources and the evaluations of the capability of using deep neural networks for temporal remote sensing analysis. Outcomes of the study shows quite impressive results over 97% of accuracy in terms of cultivated paddy area detection using optical remote sensing imagery. Moreover, the research was extended to identify cultivated paddy areas using synthetic aperture radar (SAR) imagery. It also outputs a promising result over 96% of accuracy in terms of detecting cultivated paddy regions. The study then extends to detect Brown Planthopper attacks in cultivated paddy fields.

Brown Planthopper is considered as the most destructive insect in paddy cultivation. There are no previous studies for identifying Brown Planthopper attacks using satellite remote sensing data under field conditions. In this study, ratio and standard difference indices derived from optical imagery are fed into a Support Vector Machine model to identify the regions affected by Brown Planthopper attacks. Using the results of cultivated paddy fields detection model as a filter, SVM model results are improved. The combined approach shows accuracy over 96% for detecting Brown Planthopper attacks.



**Keywords**: Remote sensing, Synthetic Aperture Radar, Agriculture, Rice, Deep Neural Networks, SVM, Brown Planthopper, Paddy Yield, Paddy Extent

## ACKNOWLEDGEMENT

I would first like to express my gratitude to all my supervisors; Mrs. Vishaka Nanayakkara, Dr. Amal Shehan Perera, Mr. Suranga Jayasena and Mr. Lasantha Fernando who always shared their wisdom and perception to make this research study a success. They constantly steered me in right direction whenever they found I needed it. Further, I would like to thank Dr.Harsha De Silva, honorable state minister of National Policies and Economic Affairs who emphasized practical challenges in policy making during a discussion which inspired me to conduct this study. I appreciate the support I received from the Ministry of National Policies and Economic Affairs and the Ministry of Agriculture during the implementation of my research.

I would also like to acknowledge all the field officers and other officials attached to Ministry of Agriculture for committing their precious time towards this project by supporting me to collect ground data and sharing their valuable insights regarding paddy cultivation.

I am sincerely grateful for Prof. Sanath Jayasena and Dr. Supunmali Ahangama for reviewing this research study at progress reviews and providing their valuable suggestions to improve the project outcomes.

I would also like to thank Mr. Kumaran Kuganathan for reviewing the drafts of research papers and helping me to collect ground data without any hesitation. Further, I thank both academic and nonacademic staff of Department of Computer Science and Engineering for giving their fullest support towards this project.

Last but not least, I would like to express my gratitude to all the members of DataSEARCH center of Department of Computer Science and Engineering for their constant monitoring and invaluable suggestions to make this study a success.

# TABLE OF CONTENT

# LIST OF FIGURES

## LIST OF TABLES

## LIST OF ABBREVIATIONS

| Abbreviation | Description |
|---|---|
| SAR | Synthetic Aperture Radar |
| BPH | Brown Planthopper |
| NDVI | Normalized Difference Vegetation Index |
| LSWI | Land Surface Water Index |
| NDWI | Normalized Difference Water Index |
| RI | Ratio Index |
| SDI | Standard Difference Index |
| CNN | Convolution Neural Network |
| LSTM | Long Short Term Memory |
| SVM | Support Vector Machine |
| STARFM | Spatial and Temporal Adaptive Reflectance Fusion Model |
| MODIS | Moderate Resolution Imaging Spectroradiometer |

# 1. INTRODUCTION

Rice is considered as one of the most important cereal grain which is consumed as staple food by large portion of the world's population, particularly in Asian countries. Farmers in the countries of Asia pacific region including Sri Lanka is accountable for over 90% of the world's rice production which accounts to more than 750 million tons [1]. Paddy fields occupy 34% of total agricultural crop land areas in Sri Lanka [2]. Moreover, about 1.8 million farming families are engaged in rice cultivation in Sri Lanka [2]. Enhancing the productivity of farmers have been identified as one of the foremost approaches of eradicating poverty and food security but nevertheless, flawed decisions deceived by erroneous information cause the productivity to diminish.

Stakeholders attached to paddy cultivation such as farmers, government, traders, etc. require precise and timely quantitative estimations of cultivated paddy extent and yield in order to make accurate decisions on various occasions such as planting, calculating, buying, importing, exporting, etc.

On the other hand, rice cultivation is often affected by a variety of diseases and pest attacks. Hence, precise estimations regarding pests and disease attacks on paddy cultivation is vital for stakeholders to take accurate decisions on various aspects such as insurance premium calculation, evolvement of pests and disease management systems, etc. Moreover, it's also important to distinguish damaged paddy regions and healthy paddy regions in order to provide more accurate rice production estimations. In terms of pests and diseases detection, we mainly focus on detecting Brown Planthopper attacks in this study since it is the most devastating pest disease that causes paddy yield losses [3].

In Sri Lanka, Ministry of Agriculture and Department of Census and Statistics produce estimates rice production, cultivated paddy extent and the damages caused by pests and diseases attacks [4]. Those existing approaches of getting estimations heavily depend on the human involvement. Government of Sri Lanka deploys field officers to collect data by monitoring paddy cultivation around the country. Each officer sends the estimation of yield, cultivated paddy extent and damaged areas

within the assigned region to a central authority to calculate the total productive cultivated extent and other analysis purposes. Authorities have been following said conventional protocol to estimate rice production and productive paddy extent for decades. Inefficiency and inaccuracy of collecting data on the production, cultivated paddy extent and damaged areas from disease and pest attacks make paddy cultivation related decision making challenging. For instance, field officers rely on their intuitions rather than actual observations when measuring the damaged paddy areas, since either measuring small patches of pests and disease attacks is laborious or farmers are reluctant to share their information on small pest and disease attacks in their fields to field officers because no compensation is granted by the government for small-scale yield losses. Burden of handling human resources, the expenditures on human resources, and the time consumption have been identified as other major limitations of the existing approaches.

### 1.1 Remote Sensing

Stakeholders, particularly the government of Sri Lanka is eager to replace the current approach of estimating productive paddy extent with an efficient alternative mechanism. Researchers have demonstrated the capability of remote sensing data as a potential alternative for providing valuable insights about various aspects of agriculture [5] [6] [7]. Remote sensing based energy reflectance measures of various electromagnetic wavelengths from the earth's surfaces can be used to assess the condition of the plants [8] and distinguish crops from other objects on the surface [9]. Remote sensing data can be categorized into two types as follows.

- Active remote sensing
- Passive remote sensing

Active remote sensing data are collected by emitting electromagnetic energy towards the earth surface (e.g.: Synthetic Aperture Radar (SAR)) and measuring the energy that is reflected back or backscattered from the earth surface and objects on earth surface. Passive remote sensing data are collected by measuring natural electromagnetic energy that is emitted or reflected by the earth surface and objects.

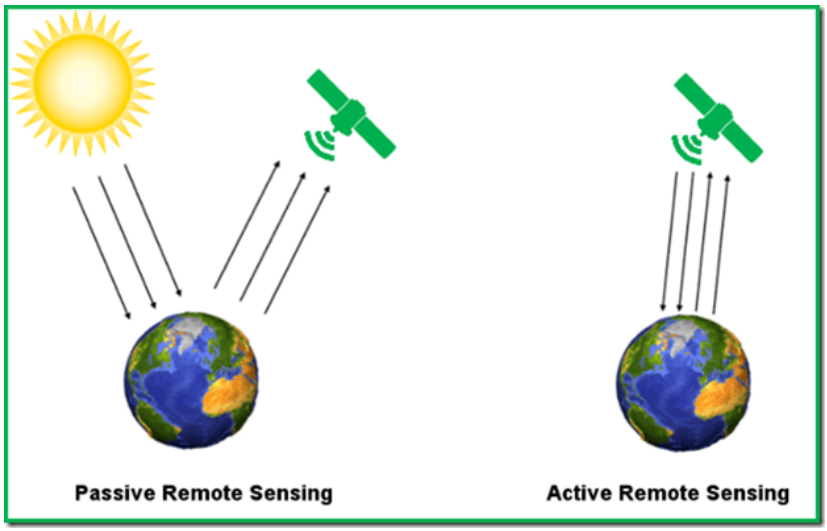


*Figure 1.1 Active vs Passive Remote Sensing*
*Source: https://grindgis.com/remote-sensing/active-and-passive-remote-sensing*

Compared to active remote sensing data, passive remote sensing data have been extensively used in past research studies due to the constraints of active remote sensing data such as limited availability, inconsistency, complex data structures, etc. [10]. With the launch of Sentinel 1, remote sensing research community now has access to SAR imagery at the scale that is needed for successful agricultural monitoring applications. On the other hand, there are limitations of using conventional passive remote sensing data resources for examining and analyzing cloud-prone and scattered agricultural regions, particularly during rainfall onsets. Emergence of novel passive remote sensing data resources with high temporal and spatial resolution, enables researchers to broaden their research studies into the areas where fine-grained information is required.

## 1.2 Motivation and Objectives

### 1.2.1 Motivation

From this background, it is clear that rice is the most economically important crop that impacts a vast majority of households in Sri Lanka. Decisions made by the government based on erroneous information, have been identified as the root cause of the recent economic impacts related to paddy cultivation. Consequently, the motivation of this study is to inform decision making in terms of productive cultivated paddy extent in a given cultivation season.

### 1.2.2 Aim and objectives

The aim of this research is to develop a methodology for estimating cultivated paddy extent and the BPH attack prevalence using remote sensing data and machine learning techniques. To achieve the aim, the following objectives are set.

- Identify remote sensing data which can be used to estimate cultivated paddy extent.
- Identify key relationships between paddy cultivation and remote sensing data.
- Implement and validate different machine learning models to identify cultivated paddy fields.
- Evaluate machine learning techniques to detect cultivated paddy extent.
- Identify remote sensing data that can be used to detect BPH attacks.
- Identify key relationships between BPH attacks and remote sensing data.
- Implement and validate machine learning model to identify areas affected by BPH attacks.

The study consists of two phases. First phase identifies the cultivated paddy field areas. In the first phase, the applicability of both active remote sensing and passive remote sensing data are tested in terms of identifying cultivated paddy fields. We use PlanetScope satellite constellation as one of the passive remote sensing data resource which consists of approximately 120 satellites providing images with 3m spatial resolution and one-day temporal resolution [11]. We use Sentinel-2, developed by European Space Agency (ESA), as the second passive remote sensing data resource. Sentinel-2 provides remote sensing data with 10m of spatial resolution and 5-days of temporal resolution [12]. Pixel-based classification models are built to evaluate the performances of each passive remote sensing data resource separately as well as the hybrid usage of the two remote sensing data resources. Particularly, we show how the proposed hybrid approach minimize the constraints of each of the remote sensing data resources. Further, the performances of two neural networks: Long Short Term Memory Network (LSTM) and Convolution Neural Network (CNN), are compared with Support Vector Machine (SVM) model as the baseline in terms of the distinguishing of cultivated paddy lands.

Sentinel 1 SAR acquisitions are used as the active remote sensing data resources in the first phase. The first Sentinel 1 satellite; (1A) and the second Sentinel 1 satellite; 1B were launched in 2014 and 2016 respectively [13]. It provides radar remote sensing data at 10m spatial resolution and 12-day temporal resolution. Pixel-based, GLCM variance time series are classified using a convolution neural network to identify cultivated paddy fields. Then the capabilities and the limitations of using passive and active remote sensing data in terms of identifying cultivated paddy fields are discussed thoroughly.

Further, the areas affected by BPH attacks within the cultivated paddy fields are identified in the second phase of the study. PlanetScope remote sensing data are used to identify the areas damaged by Brown Planthopper attacks. A Support Vector Machine model based on ratio and standard difference indices is developed to identify the affected areas by BPH attacks.

Authorities have defined average rice yield productions per hectare for different regions in the country. Hence, Phase 1 and Phase 2 calculations of cultivated paddy extent and BPH attacks then can be used to derive the paddy yield estimations in future steps.

Chapter 2 discusses the remote sensing resources and deep learning techniques used in the study. Further, it dives into the past research work conducted with regard to crop classification and pests and disease detection. Chapter 3 explains two types of approaches for estimating cultivated paddy extent. The approach to detect Brown Planthopper attack using remote sensing and machine learning techniques is discussed in Chapter 4. A comprehensive discussion regarding Agri AI platform and the conclusion are included in Chapter 5 and Chapter 6 respectively.

# 2. BACKGROUND AND RELATED WORK

In this section, we discuss the prior works conducted in the areas of crop land classifications and disease detection. We look at applicable technologies and remote sensing data resources used, for this study. Further, we discuss the prior studies related to our study and what technologies, data resources and methodologies are used in previous research work.

## 2.1 Vegetation Indices

Vegetation indices are largely used for agricultural analysis operations such as crop land classification, crop yield prediction and crop disease detection etc. [14] [15] [16]

### 2.1.1 Normalized Difference Vegetation Index

A substantial number of studies have proven that Normalized Difference Vegetation Index (NDVI) has a significant efficacy in classifying crop lands [17] [18]. Leaves absorb energy from sunlight in certain wavelengths of electromagnetic spectrum while the remaining wavelengths are reflected back. Chlorophyll residing in leaves absorb energy from red spectral band in visible light for photosynthesis process while cell structure of leaves reflects energy in near-infrared spectral band. NDVI is derived by measuring reflectance of red (RED) and near-infrared (NIR) spectral bands as shown in the formula 1 below. Values close to zero indicates no healthy green leaves and values close to +1 indicates high density of healthy green leaves.

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)} \quad (1)$$

A vast number of research studies have exploited the potential of MODIS and LANDSAT NDVI data in terms of classifying crop lands in large agricultural lands [17] [18] [19] [20] [21] such as US Central Great Plains, Great Lakes Basin, etc. They were able to distinguish the land into a variety of crop types with decent accuracies using various machine learning algorithms ranging from simple decision tree classifiers to deep neural networks.

### 2.1.2 Land Surface Water Index

During the initial growing period of flooding and transplanting, water level in paddy field surface increases substantially compared to other crop lands. This discrete characteristic of paddy lands during the initial growing stage can be observed by Land Surface Water Index (LSWI) since it is highly responsive to the total amount of liquid water in vegetation and its soil surface [22]. Shortwave infrared (SWIR) and the near infrared (NIR) wavelengths of electromagnetic spectrum are used to derive LSWI as in formula 2. In fact, the index derived using the NIR and SWIR wavelengths has different nomenclatures in different research. In few, it was referred as Normalized Difference Water Index (NDWI) instead of Land Surface Water Index (LSWI) [23] [24]. In this study, it is referred to as LSWI.

$$LSWI = \frac{(NIR - SWIR)}{(NIR + SWIR)} \tag{2}$$

There were previous efforts of combining LSWI with other vegetation index parameters in crop land classification tasks [25] [26]. Xiangming Xiao et al. [25] developed an algorithm to detect cultivated paddy fields incorporating three types of vegetation indices; NDVI, EVI and LSWI. Lv Tingting et al. [26] used LSWI along with several other vegetation index parameters to understand cropping patterns, crop intensity and crop land mapping.

## 2.2 Neural Networks

In this decade, deep neural networks have gained a wide-spread attention, mainly by outperforming other conventional machine learning approaches in a broad area of applications, especially in image processing and natural language processing [27]. There are two main reasons behind the recent success of deep neural network applications:

- Availability of massive data sets
- Rapid technological development of hardware.

### 2.2.1 Feed Forward Neural Networks

Feed-forward Neural Network architectures are the most versatile deep neural network types which clearly predominate conventional machine learning algorithms [28]. Those architectures are based on strict sequential data propagation without forming cycles between nodes. Feed-forward neural network consists of connected neurons (Figure 2.1 (a) and Figure 2.1 (b)) which activates its output conforming to Formula 3.

$$y = \sigma(\sum_{i} W_{ij} x_i + b) \tag{3}$$

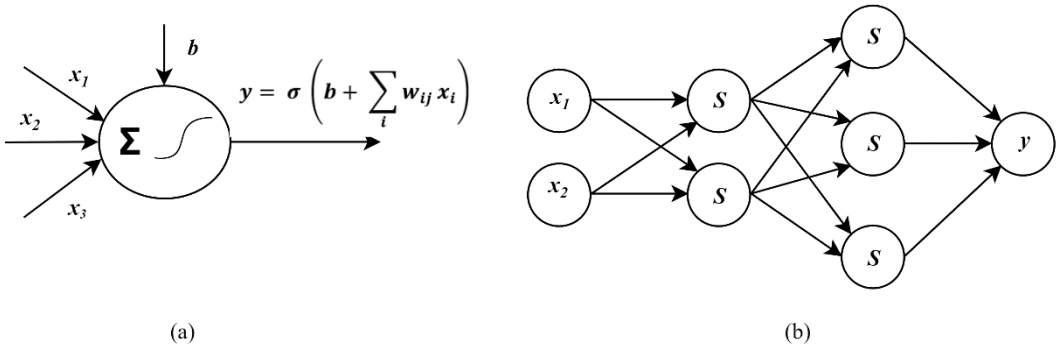


*Figure 2.1 Architecture of Feed Forward Neural Network. Figure 2.1(a) shows a perceptron. Figure 2.1(b) shows a feed forward neural network which consists of multiple layers of perceptrons*

### 2.2.2 Recurrent Neural Networks

Feed-forward neural network architectures are capable of learning from individual uncorrelated data items, but not favorable to process time-dependent sequence data since it requires remembering previously processed data items. In contrast, Recurrent Neural Networks solve the issue of sequence handling to a reasonable extent, but not perfectly. Edges that connect adjacent time steps form cycles, including cycles of length one that are self-connections from a node to itself across time [29]. Formula 4 and Formula 5 show the necessary computation at each time step in a simple recurrent neural network shown in Figure 2.2.

$$a<t> = g(w_{aa} a<t-1> + w_{ax} x<t> + b_a) \tag{4}$$

$$y<t> = g(w_{ya} a<t> + b_y) \tag{5}$$

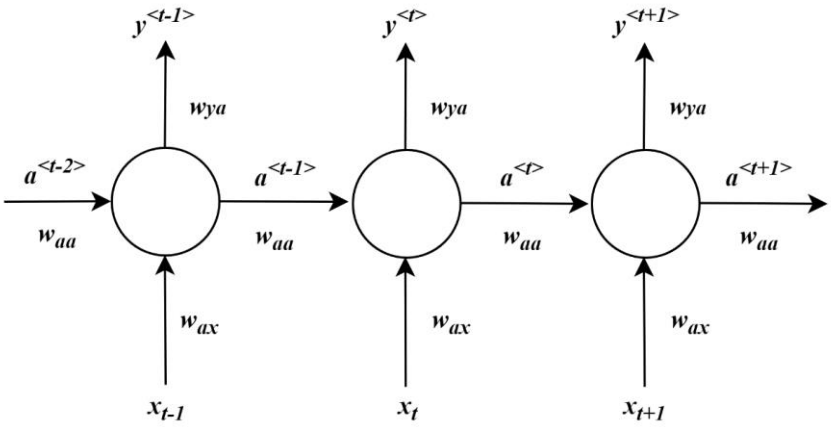


*Figure 2.2 Architecture of a simple Recurrent Neural Network*

At time t, a<t> represents the values of the current hidden node while a<t-1> represents values of the previous state's hidden node. $b_y$ and $b_a$ are bias parameters which allow each node to learn an offset. x<t> denotes the current data point. $w_{aa}$, $w_{ax}$ and $w_{ya}$ are the corresponding weights applied to a<t-1>, x<t> and a<t>.

Recurrent Neural Networks turn out to be quite effective when we are dealing with short-term dependencies. At each time step, the RNN must encode information it has received from previous time step and pass the new information through a set of feedback connections to next time step. It is very challenging to prevent the information degradation over a long time sequence. RNNs are also affected by vanishing gradient, preventing back-propagating error correcting information through the model, losing the connection to long history as time progress. Hence, RNN may fail to grasp long term transitional patterns in paddy cultivation.

### 2.2.3 Long Short-Term Memory Network

LSTM network consist of LSTM cells which maintain a state called cell state (c<t>). It enables a way to selectively remember and forget long-term information. The information of a particular cell state depends on three aspects:

- previous cell state (c<t-1>)
- previous hidden state (a<t-1>)
- input at current time step (x<t>)

Three gates; forget gate, input gate and output gate, are derived using the previous hidden state (a<t-1>) and current input (x<t>) as shown in the Figure 2.3. Current

cell state either remains nearly similar to previous cell state value or gets updated, based on the forget and update gate values. Current hidden state value is derived based on the value of the output gate and the current cell state (c<t>). Formula 6,7,8,9,10 and 11 are used to calculate the corresponding values of the LSTM cell.

$$f_t = \sigma_f(W^f_{data}x_t + W^f_{state}a_{t-1} + b^f) \quad (6)$$

$$i_t = \sigma_i(W^i_{data}x_t + W^i_{state}a_{t-1} + b^i) \quad (7)$$

$$\hat{c}_t = \sigma_{\hat{c}}(W^{\hat{c}}_{data}x_t + W^{\hat{c}}_{state}a_{t-1} + b^{\hat{c}}) \quad (8)$$

$$o_t = \sigma_o(W^o_{data}x_t + W^o_{state}a_{t-1} + b^o) \quad (9)$$

$$c_t = f_t * c_{t-1} + i_t \odot \hat{c}_t \quad (10)$$

$$a_t = o_t \odot \sigma_a(c_t) \quad (11)$$

M. Rußwurm et al. [28] conducted a study using remote sensing data captured by Sentinel 2A to emphasize the importance of multi-temporal approaches over mono-temporal approaches for land cover classification. They have employed long short-term memory neural network for crop identification purposes. Further, the performance was compared with a mono-temporal convolution neural network and support vector machine model. An approach based on LSTM has been presented for land cover classification in Florida Everglades ecosystem study site. Reported patch based multi-image LSTM model outperformed pixel based single image NN significantly with over 30% of accuracy improvement.

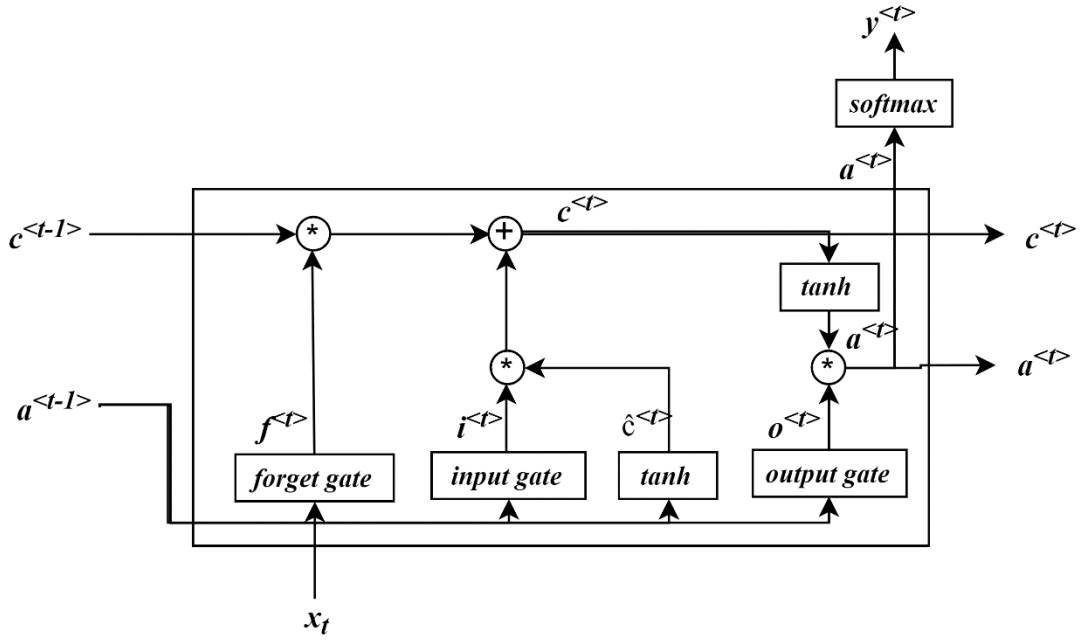


*Figure 2.3 Architecture of a LSTM unit cell*

### 2.2.4 Convolution Neural Networks

Architecture of 1D CNN is shown in Figure 2.4. Typically, it is a composite of multiple convolution, pooling, activation, fully connected layers. The convolution layer output a feature map by calculating dot product between an input feature map and a set of weights [30]. The feature map is commonly known as filters. In the pooling layer, size of the feature map will get reduced to downsize the number of parameters and computation in the network. Nonlinear activation functions are leveraged in between convolution layers. Relu activation function can be considered as the most common activation function used in studies. The fully connected layer and Softmax activation layer are stacked at the end of the network to calculate the probability scores for each defined class. Neural networks are often over fitted to the training data set and perform poorly in test data set due to various reasons such as a smaller number of data points in the training data set. There are several regularization techniques followed in machine learning paradigm in order to overcome the issue of overfitting. Dropout was recently introduced by Srivastava et al. as a way of regularization for deep neural networks [31]. Input data are fed through the neural network in a feedforward manner while it gets trained using

stochastic gradient descent (SGD) extensions such as Adam Optimizer in a manner of backpropagation.

Influences of convolution neural networks are extending boundaries of the remote sensing research community, particularly to classify crop lands using high resolution remote sensing imagery [32]. Moreover, recent studies have revealed that one dimensional CNNs are suitable for various types of time series classification applications [33]. However, one dimensional convolution neural networks (1D CNN) have been very rarely used for remote sensing time series classification.

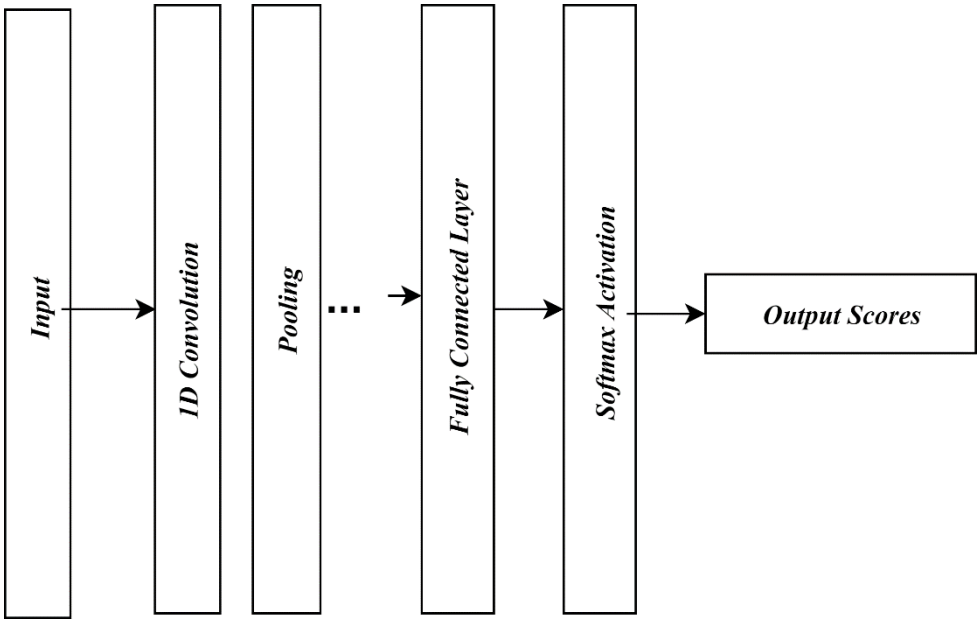


*Figure 2.4 A simple archtecture of a 1D CNN*

### 2.2.5 Support Vector Machines

Support Vector Machine (SVM) model is a supervised machine learning model for classification tasks and regression analysis. In SVM approach, data points are mapped into a space and divide the space into classes. Then the test data points are assigned to the space to determine its class. Even though it was originally designed for binary linear classification, it can be also used for non-linear multi class classification using a methodology called kernel.

Support Vector Machine models have been widely used in agricultural crop mapping tasks [34] [35]. For instance, Baojuan et al. [36] developed a SVM model which can identify nine crop types in a complex agricultural management system using Landsat

NDVI datasets with an accuracy of over 86%. A comparison conducted by Yang et al. reveals that Support Vector Machine models performs better compared to neural networks in crop mapping applications when the dataset is small [37]. Moreover, Support Vector Machine has showed promising capabilities to identify various crop diseases using hyper spectral analysis [38] [39]. Liu et al. [39] has demonstrated the potential of support vector machine to discriminate two paddy diseases and healthy plants using visible and NIR spectral band data.

### 2.3 Passive Remote Sensing Data with Low Spatial and Temporal Resolution for Crop Land Classification

Landsat and MODIS remote sensing data are being extensively used in studies related to mapping agricultural croplands [17] [18] [19] [20] [21], but these resources have key limitations when the accuracy is important. In most cases, one or two, 250m MODIS satellite image pixels are enough to cover an entire crop land in Sri Lanka. Thus, it would be challenging to derive meaningful and accurate insights regarding agricultural productivity [40]. Linear mixture models have been introduced to resolve issues associated with coarser resolution remote sensing data for crop land mapping in areas where mixed cropping agriculture are practiced [41].

Despite having comparably high spatial resolution than MODIS satellites, Landsat satellites' low temporal resolution limit their capability of being used as a sole data source for agricultural analysis purposes. Cloud contamination further constraint the usability of LANDSAT in cloud-prone areas, particularly in wet zones. There were significant number of attempts of fusing MODIS and Landsat vegetation data in order to mitigate the limitations posed by individual data sources [18] [35] [37]. Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) [42] was used by researchers frequently to fuse Landsat and MODIS data in crop classification applications [32] [34].

### 2.4 Passive Remote Sensing Data with High/Medium Spatial and Temporal Resolution for Crop Land Classification

Until recent years, freely available remote sensing data resources were not suitable enough to monitor small crop lands in more granular level due to low spatial and temporal resolutions [43] [14]. Recently, inexpensive high and medium spatial and

temporal imagery have been widely available due to the recent emergence of finer technologies and involvement of private institutes that own satellite constellations. On the other hand, satellites which offer high spatial resolution, cover a smaller size of area footprint per scene. It greatly increases the amount of image processing tasks such as geometric registration, radiometric normalization and classification, etc. Advances in the field of high performance computing has reduced the burden of said image processing tasks.

With the launch of Sentinel-2, researchers have started to investigate its suitability for agricultural applications. Markus Immitzer et al. [44] conducted a study based on pre-operational Sentinel-2 data for classifying crop types and tree species in Lower Austria. Marshall et al. discussed the capability of intensive field sampling 1-m Terra Bella imagery to monitor individual smallholder farms in Kenya [40]. Further, their comparison among different remote sensing data resources reveals that explanatory power at 10m is approximately only about 75% of what it is at 1m and similarly at 30m resolution explanatory power falls by approximately 50%.

### 2.5 Active Remote Sensing Data for Crop Land Classification

There are regions in Sri Lanka where a single clear optical remote sensing image during monsoon season is unavailable due to heavy cloud contamination. High temporal passive remote sensing data can only mitigate the effects of cloud contamination to a certain extent. Synthetic Aperture Radar (SAR) imagery allows us to monitor crop lands and classify them without the effects of cloud contamination [45]. Substantial a number of studies have been conducted by combining SAR imagery with optical imagery for crop mapping [45] [46]. In some cases, crop classifications based on SAR imagery outperforms crop classifications based on optical imagery [46]. Moreover, combined crop classification approaches using SAR and optical imagery reported better accuracy compared to crop classification approaches based only on optical imagery [45] [46]. Kun Jia et al. [47] reveal that crop mapping based on multi-configuration (using multiple SAR datasets) is more accurate than classifications based on single SAR image with single configuration. They also stated that exploiting texture features of SAR imagery (e.g. contrast, correlation, energy) enhances the accuracy of crop mapping [47]. Further, the

temporal variations of radar reflectance have been investigated as a way of crop land mapping.

### 2.6 Paddy Disease Detection

There are only a few attempts to use remote sensing for paddy disease detection. Studies based on satellite remote sensing are even more rare in the area of paddy disease detection due to limited access to satellite remote sensing data resources with high temporal and spatial resolution until recent years. Most of the research attempts related to paddy disease detection, have been relied on ground based sensors or airborne remote sensing data resources [48]. When a plant is affected by diseases, often its reflection in visible and infrared region deviates from natural reflection due to internal tissue structure changes, decline of the chlorophyll content and other pigment changes [49]. Hence, various spectral ratio indices, vegetation indices and standard difference indices have been proposed to identify crop diseases [42] [50]. Many researchers have observed a drastic drop of reflectance in near infrared band and red band in plants affected from various stresses including diseases [51]. Since the past studies relied on ground sensors and airborne remote sensing data with high spatial and temporal resolution, researchers were able to distinguish even the severity of the paddy diseases. Instead of using various indices, Z.-Y. Liu et al. used raw spectral reflectance (from 400nm to 2400nm) and its variations (e.g. second and second derivatives) to detect fungal infections at laboratory conditions [52]. Moreover, Z. Yong et al. showed that it is difficult to distinguish two different wheat stresses by only analyzing single spectral band [53]. In contrast, it's possible to discriminate two different wheat stresses by making use of appropriate vegetation indices. But it suggests that using the models that were built to identify plant stresses from canopy level observations is challenging to up-scale into the field level due to dynamic conditions of agricultural fields [53].

Only few studies have been conducted to detect Brown Planthopper attacks using hyper-spectral reflectance measurements. Chwen-Min et al. has identified the hyper-spectral characteristics (both single spectral bands and different combinations of spectral bands) associated with identifying BPH attacks and severity assessments of BPH attacks [54]. Moreover, they added that their study shows a significant

capability of using appropriate spectral bands to distinguish BPH attacks and Leaffolder attacks. Prasannakumar et al. proposed a multiple linear regression model which incorporates field level measurements of hyper-spectral reflectance to identify BPH damage level [55]. They suggested to use normalized ratios of band reflectance to mitigate the consequences of externalities such as solar angle, soil background and crop growth [55].

# 3. CULTIVATE PADDY EXTENT DETECTION

In this phase, an estimation for cultivated paddy extent is taken through two approaches: i) using Passive Remote sensing data, ii) using Active Remote Sensing data. Procedures of those two approaches as well as pros and cons of those are discussed in this chapter.

## 3.1 Cultivated Paddy Extent Detection using Passive Remote Sensing Data

In this section, a novel approach for estimating cultivated paddy extent is proposed based on two passive remote sensing data resources and deep learning. Further, we evaluate the performances of the two remote sensing data resources and LSTM, CNN neural networks in terms of detecting cultivated paddy areas.

### 3.1.1 Study area

Three agro-ecological zones can be determined in Sri Lanka namely, Dry zone, Intermediate zone and Wet zone. The land is divided into aforesaid agro-ecological zones in terms of the heterogeneity of rainfall, soil and vegetation. Regions of the island that cover western, south and hill country belongs to Wet Zone. Dry zone covers eastern and northern regions, being separated by Intermediate Zone (see Figure 3.1). Wet zone receives the highest average annual rainfall of over 2000mm whereas in Dry Zone, the annual rainfall is between 800mm and 1200mm [56]. There are four monsoon seasons in Sri Lanka:

- First Inter Monsoon rains
- South West Monsoon rains
- Second Inter Monsoon rains
- North East Monsoon rains

which bring diverse rainfall regimes resulting high agro-ecological heterogeneity even though Sri Lanka only have a small aerial extent. As mentioned previously, there are two major paddy cultivation seasons namely:

- Maha Season
- Yala Season

The most common practice is starting the Maha season with the rainfall onset in Mid-September or early October and continuing till late January or early February. In general, Yala season starts from mid of March and continues up to early May. Yala season is considered as a minor cultivation season as the rainfall during Yala season is less than the rainfall in Maha Season. On the other hand, farmers in dry zone mainly rely on irrigated cultivation than rain-fed cultivation, as dry zone receives considerably low precipitation. However, the farmers in Wet Zone and Intermediate Zone are heavily depend on rain-fed cultivation than irrigated cultivation. Size of an individual landscape element (paddy field) is below 50 acres on average in Wet Zone. In contrast, individual landscape element covers over 60 Acres of aerial extent in Dry Zone.

We picked a set of sample paddy fields covering all three agro-ecological zones to conduct our research. Sample paddy fields in Wet Zone are picked from Baddgeama (See Figure 3.1) covering 80 acres. Sample paddy fields in Dry Zone are picked from Anuradhapura (See Figure 3.1) covering 120 acres. Sample paddy fields in Intermediate Zone are situated at Polgahawela covering the extent of 100 acres as showed in Figure 3.1.

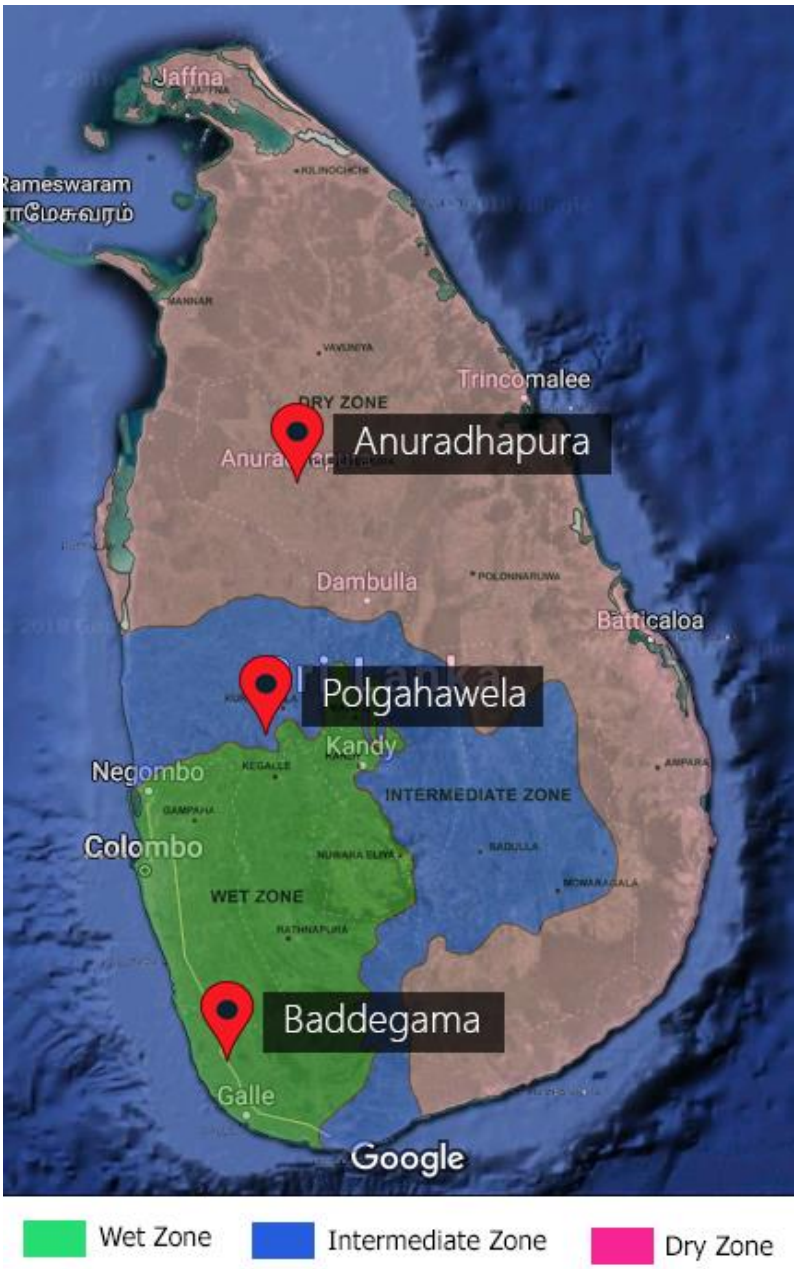


*Figure 3.1 Agro-ecological regions and locations of sample paddy fields. Samples are collected from Baddegama(Wet zone), Polgahawela(Intermediate zone) and Anuradhapura(Dry zone)*

### 3.1.2 Study data

This section of our study uses multi-spectral time series remote sensing data gathered from PlanetScope and Sentinel-2. Remote sensing data related to 2017-2018 Maha season that was started approximately from Mid-September or early October (varied with rainfall onset) are used to derive LSWI and NDVI time series. Obtaining remote sensing data which were not affected by cloud contamination is difficult despite the

fact that PlanetScope has one day temporal resolution (1-day revisit time). At least 17 PlanetScope remote sensing scenes are used to obtain time series of each sampled pixel. Since PlanetScope only collects reflectance of four spectral bands; Red, Green, Blue and Near Infra-red (NIR), it is only possible to derive NDVI time series from PlanetScope dataset. Further, acquiring remote sensing data from Sentinel-2 is more challenging since it has a longer revisit time of 5 days. Despite having longer revisit time, Sentinel-2 captures reflectance of 13 spectral bands. Thus, both NDVI and LSWI time series can be derived from Sentinel-2 spectral band data set. At least 9 Sentinel-2 remote sensing scenes were used to obtain time series of each sampled pixel.

Sampled pixels are labeled corresponding to the ground truth data collected by the field officers. Training and test data sets are selected with care in a way that both data sets include data points from every agro-ecological zone in Sri Lanka. Moreover, training and test data sets are picked from separate paddy fields in order to maintain the independence of the data sets. Data are extracted covering ten and fifteen paddy fields in each agro-ecological zone from PlanetScope and Sentinel-2 respectively. 200 non-adjacent pixels in each PlanetScope scene and corresponding re-sampled 200 pixels in each Sentinel 2 scene are annotated as cultivated or non-cultivated from each selected paddy field in order to construct models that use PlanetScope data or both PlanetScope and Sentinel-2 data. Another 1000 non-adjacent, non-cultivated pixels are selected from the non-paddy fields areas in each agro-ecological zone. Moreover, 200 non-adjacent, non-re-sampled Sentinel-2 pixels are annotated from each paddy field in order to construct machine learning models which only use Sentinel-2 data. Further, 1000 of another non-adjacent, non-cultivated and non-re-sampled pixels are labeled from the areas other than paddy fields for those models. Consequently, each of the three classification approaches is developed based on 9000 independent data points. Pixels which cover the boundaries of paddy fields are not included in the selected data sets.

### 3.1.3 Preprocessing steps

PlanetScope data products are subjected to several preprocessing steps prior to releasing the products for general analytical tasks. There are several analytical ready PlanetScope product types that are categorized according to the undergone preprocessing steps. Ortho tile product type is used in this research since those products have been corrected for geometric, radio-metric, terrain and sensor distortion errors. PlanetScope Ortho tile scene consists of four spectral bands and we use third (red) and fourth (NIR) spectral bands to derive Normalized Difference Vegetation Index (NDVI) (See Equation 1 of page 6). Every PlanetScope remote sensing scene is released along with unusable data mask file. It indicates the areas that consist of corrupted pixel values due to various reasons (e.g. cloud contamination).

Similarly, we used Level-1C Sentinel-2 data products in this study. We use fourth (Red) and eighth (NIR) spectral bands to derive NDVI and eighth (NIR) and eleventh (SWIR) spectral bands to calculate LSWI values (See Equation 2 of page 7). Level-1C data comprises quality indication files which indicate the areas affected by cloud contamination. We use those quality files to avoid considering damaged pixel values from the analysis.

Normalized Difference Vegetation Index (NDVI) values are calculated for all the pixels in each PlanetScope image. Then we use the unusable data mask file to identify and mark damaged pixels in each PlanetScope image. Analysis of some of the selected paddy fields is required merging two Ortho tile PlanetScope images captured in the same day as the aerial extent of those paddy fields are split into two Ortho tile images. Final step of preprocessing is deriving time series for each data point (pixel). Pixel values that are marked as damaged (e.g.: affected by cloud contamination) are excluded from time series. Continuous time series of selected pixels are derived by applying linear interpolation to fill the missing values. We keep temporal granularity of the time series as one week as changes of vegetation indices within the time period of one week provide sufficient analysis capabilities. Figure 3.2 shows two PlanetScope NDVI time series demonstrating the time series

characteristics of a cultivated paddy pixel (Figure 3.2(a)) and an abandoned paddy field pixel (Figure 3.2(b)). Selection procedure of training and test data sets were carefully carried out in a way that time series of a selected pixel has a least effect of low quality pixel values.

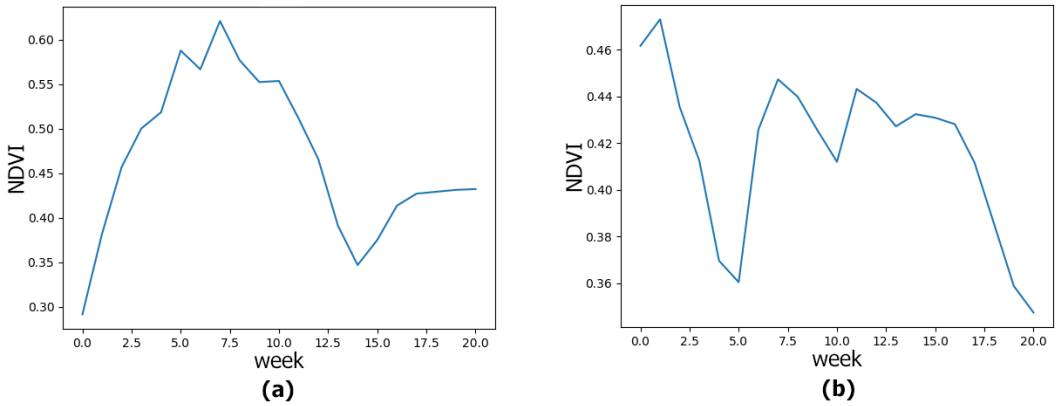


*Figure 3.2 PlanetScope NDVI time series graphs*

Similarly, as in PlanetScope preprocessing procedure, merging and cropping steps are applied for Sentinel-2 remote sensing data. Both LSWI and NDVI vegetation indices are derived from Sentinel-2 spectral bands. Each selected pixel from PlanetScope imagery should have a corresponding pixel in Sentinel-2 imagery with similar Geo-coordinates. Thus, we converted 10m resolution Sentinel-2 pixels into 3m resolution pixels by resampling. LSWI time series of each pixel is then derived by applying linear interpolation to fill missing values. We picked another set of pixels without re-sampling to derive LSWI and NDVI time series from Sentinel-2 imagery. Every low quality pixel value is eliminated prior to deriving NDVI and LSWI time series. Figure 3.3 shows two Sentinel-2 NDVI time series showing the NDVI time series characteristics of cultivated paddy pixel and abandoned paddy field pixel. Moreover, Figure 3.4 exhibits Sentinel-2 LSWI time series of a cultivated paddy field and abandoned paddy field pixel. The complete preprocessing procedure is depicted in Figure 3.5.

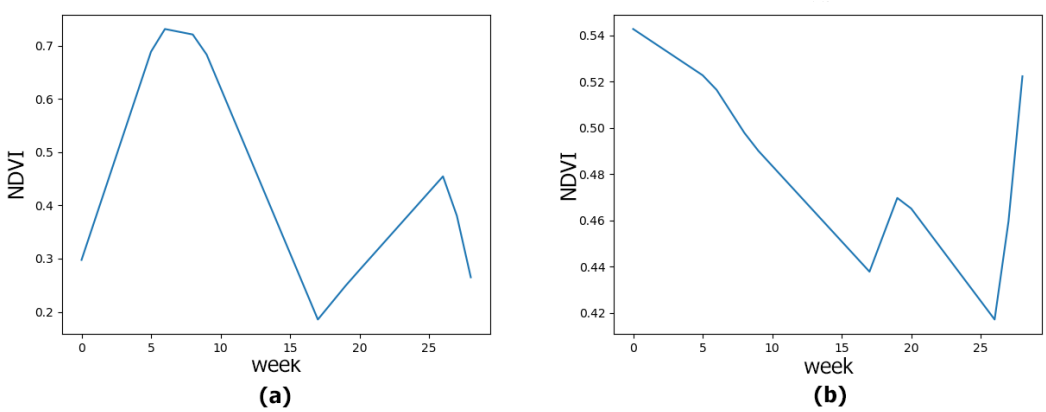


*Figure 3.3 Sentinel-2 NDVI time series graphs*

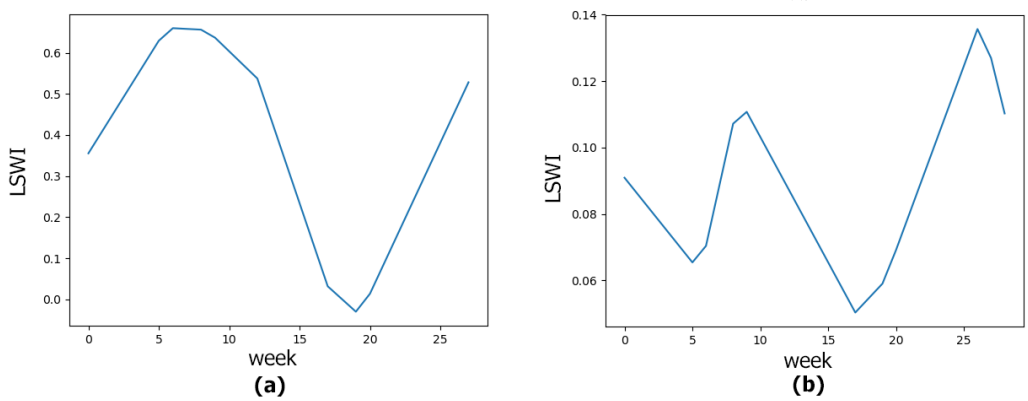


*Figure 3.4 Sentinel-2 LSWI time series graphs*

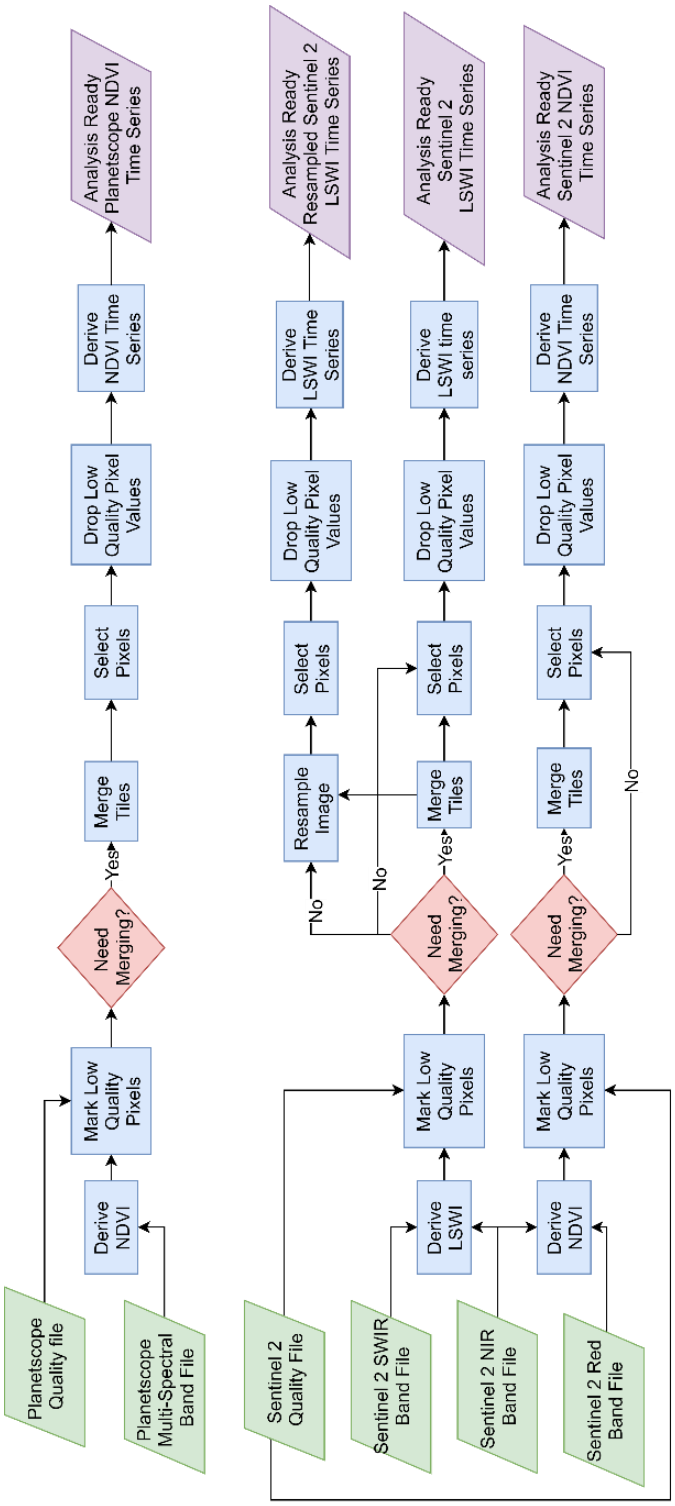


*Figure 3.5 PlanetScope and Sentinel 2 data Preprocessing Procedure*

### 3.1.4 Data partitioning

As previously mentioned, we separate the data set into two small data sets as training data set and testing data set. Following three aspects are taken into consideration when assigning data to small data sets.

- Equal class distribution between training and testing data sets.
- Equal data point distribution among data sets in terms of agro-ecological zones.
- Independence of data sets.

We shuffle entire data set and then randomly assign data points to each data sets manually in order to divide data points between data sets conforming to said criteria. Training data set consists of 8000 data points and test data set consists of 1000 data points. We follow 10 fold cross validation for training and validation process of neural networks. Hence, we further divide the selected 8000 training set into 10 smaller data chunks of 800 data points. Each neural network is run for 10 iteration using a different chunk as validation data set in each iteration and remaining 9 chunks are used for training. This method greatly reduces the possibility of over-fitting.

### 3.1.5 Experiment setup

In this study, we compare not only the capabilities of two remote sensing resources (PlanetScope and Sentinel-2) and their hybrid approach, but also the potential of two neural networks (CNN and LSTM) in terms of classifying paddy lands. Results presented in following sections are reported by running each neural network model on test data set. Keras; [57] an open source neural network library, is used to implement all neural networks that are to be mentioned in the following sections. We run Keras neural network models on top of Tensorflow; an open source framework for machine learning developed by Google [58].

#### 3.1.5.1 Paddy land classification using NDVI

We implemented six types of classification models using NDVI time series data for evaluating the potential of PlanetScope and Sentinel-2 data sources. Further, we

evaluate LSTM and CNN models using Support Vector Machine algorithm as the baseline:

- LSTM models using PlanetScope data
- CNN models using PlanetScope data
- SVM model using PlanetScope data
- LSTM models using Sentinel-2 data
- CNN models using Sentinel-2 data
- SVM model using Sentinel-2 data

We evaluate Long Short-Term Memory neural networks and Convolution Neural Networks with various hyper parameter settings. We analyze LSTM performance with different number of layers and different number of LSTM unit cells as showed in Table 3.1. On the other hand, Convolution Neural networks are assessed with different number of layers and different number of filters as showed in Table 3.1. Every Convolution Neural Network architecture that we have discussed here, consists of a dropout layer followed by CNN layers accordingly for the sake of avoiding over-fitting. In addition to regular dropouts, recurrent dropouts are added for LSTM networks. Radial Basis Function (rbf) kernel is applied for Support Vector Machine models. Hyper parameters of Support Vector Machines are similar in every experiment setup of this study.

#### 3.1.5.2 Paddy land classification using PlanetScope NDVI and Sentinel-2 LSWI

We experiment the capability of integrating NDVI and LSWI vegetation indices in terms of identifying cultivated paddy pixels. This hybrid approach is tested with variety of neural network settings similar to above experimental setup which is used to assess the capability of PlanetScope's NDVI time series data. Compared to previous experiment setup, hyper parameter value variations of CNN and LSTM models are kept unchanged except the number of LSTM unit cells in a single LSTM layer. Results of this experimental setup are reported in Table 3.3.

#### 3.1.5.3 Paddy land classification using Sentinel-2 NDVI and Sentinel-2 LSWI

Since Sentinel-2 provides all the spectral bands that are needed to derive both NDVI and LSWI indices, we tested Sentinel-2 as the only remote sensing data resource for

paddy land classification models. Evaluations are carried out under the similar hyper parameter settings used in the previous section. Table 3.4 reports the performances of this experimental setup.

*Table 3.1 Performance Evaluation of Machine Learning Models built based on Planetscope NDVI.*

| | L(50,1) | L(75,1) | L(100,1) | L(50,2) | L(75,2) | L(100,2) | C(10,1) | C(30,1) | C(50,1) | C(10,2) | C(30,2) | C(50,2) | SVM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 91.43 | 92.52 | 87.98 | 88.66 | 85.48 | 65.19 | 57.34 | 72.26 | 84.54 | 62.12 | 77.35 | 89.27 | 98.67 |
| Recall | 91.34 | 92.49 | 88.16 | 88.7 | 85.7 | 65.6 | 58 | 72.68 | 84.59 | 59.92 | 76.79 | 88.93 | 87.45 |
| F1-score | 91.38 | 92.5 | 88.07 | 88.68 | 85.59 | 65.39 | 57.67 | 72.47 | 84.56 | 61 | 77.07 | 89.1 | 92.72 |
| Accuracy | 91.39 | 92.5 | 88.04 | 88.68 | 85.53 | 65.33 | 57.16 | 72.27 | 84.54 | 61.23 | 77.14 | 89.12 | 92.9 |

'L' stands for LSTM. First value within a bracket of a LSTM column represents number of LSTM unit cells and the second value represents number of LSTM layers.
'C' stands for CNN. First value within a bracket of a CNN column represents number of filters and the second value represents number of CNN layers.

*Table 3.2 Performance Evaluation of Machine Learning Models built based on Sentinel-2 NDVI*

| | L(50,1) | L(75,1) | L(100,1) | L(50,2) | L(75,2) | L(100,2) | C(10,1) | C(30,1) | C(50,1) | C(10,2) | C(30,2) | C(50,2) | SVM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 93.37 | 97.82 | 93.94 | 92.24 | 94.83 | 95.29 | 62.73 | 78.61 | 90.15 | 55.96 | 76.38 | 93.9 | 98.74 |
| Recall | 93.17 | 97.88 | 93.89 | 92.46 | 95.15 | 95.52 | 61.64 | 77.87 | 90.02 | 54.26 | 75.05 | 93.03 | 90.66 |
| F1-score | 93.27 | 97.85 | 93.91 | 92.35 | 94.99 | 95.4 | 62.18 | 78.24 | 90.08 | 55.1 | 75.71 | 93.46 | 94.53 |
| Accuracy | 93.28 | 97.85 | 93.92 | 92.34 | 94.97 | 95.38 | 62.48 | 78.36 | 90.07 | 55.49 | 75.93 | 93.48 | 93.27 |

Refer Table 3.1 to understand column titling.

*Table 3.3 Performance Evaluation of Machine Learning Models built based on PlanetScope NDVI and Sentinel-2 LSWI*

| | L(100,1) | L(125,1) | L(150,1) | L(100,2) | L(125,2) | L(150,2) | C(10,1) | C(30,1) | C(50,1) | C(10,2) | C(30,2) | C(50,2) | SVM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 84.1 | 80.24 | 79.95 | 89.28 | 94.87 | 96.73 | 87.75 | 92.99 | 93.2 | 87.87 | 96.58 | 96.81 | 97.04 |
| Recall | 84.02 | 79.62 | 80.41 | 88.57 | 95.5 | 96.92 | 87.85 | 93.06 | 93.23 | 87.28 | 96.67 | 96.78 | 86.99 |
| F1-score | 84.06 | 79.93 | 80.18 | 88.92 | 95.18 | 96.82 | 87.8 | 93.02 | 93.21 | 87.57 | 96.62 | 96.79 | 91.74 |
| Accuracy | 84.03 | 80.04 | 80.12 | 89.02 | 95.14 | 96.81 | 87.81 | 93.02 | 93.21 | 87.62 | 96.61 | 96.8 | 91.3 |

Refer Table 3.1 to understand column titling.

*Table 3.4 Performance Evaluation of Machine Learning Models built based on Sentinel-2 NDVI and Sentinel-2 LSWI*

| | L(100,1) | L(125,1) | L(150,1) | L(100,2) | L(125,2) | L(150,2) | C(10,1) | C(30,1) | C(50,1) | C(10,2) | C(30,2) | C(50,2) | SVM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 83.06 | 85.61 | 91.25 | 84.32 | 89.52 | 97.73 | 87.43 | 95.39 | 97.21 | 90.28 | 97.24 | 97.94 | 97.26 |
| Recall | 82.8 | 85.3 | 91.3 | 84.42 | 90.02 | 97.87 | 83.51 | 95.04 | 97.03 | 87.14 | 97.01 | 97.9 | 88.94 |
| F1-score | 82.93 | 85.45 | 91.27 | 84.37 | 89.77 | 97.8 | 85.43 | 95.21 | 97.12 | 88.68 | 97.12 | 97.92 | 92.91 |
| Accuracy | 82.97 | 85.56 | 91.27 | 84.38 | 89.72 | 97.79 | 85.88 | 95.21 | 97.12 | 88.9 | 97.13 | 97.92 | 92.35 |

Refer Table 3.1 to understand column titling.

#### 3.1.5.4 Paddy land extent estimation

The ultimate goal of this phase in the study is to suggest a sophisticated alternative methodology for estimating cultivated paddy extent. Hence, we select the best machine learning models from each of the above four classification approaches and make use of them to identify cultivated paddy lands. Three areas with larger aerial extent are selected from the three agro-ecological zones. Time series of all pixel which belong to the selected areas are derived by applying required preprocessing steps. Subsequently, time series of every pixel is classified by each selected machine learning models. Extent of cultivated paddy fields are then estimated by aggregating the area covered in each pixel that is classified as a pixel covering cultivated paddy areas. Cultivated paddy land extent from ground truth data in selected areas and estimated cultivated paddy land extent from each machine learning model are reported in the Table 3.5. Moreover, the actual cultivated paddy land extent and the estimated cultivated paddy land extent in each region are marked in Figure 3.6, Figure 3.7 and Figure 3.8.

### 3.1.6 Results

Table 3.1 shows the accuracy measurements of LSTM and CNN machine learning models with different hyper parameter settings along with the SVM model as the baseline model in terms of classifying paddy lands based on only PlanetScope NDVI data. In this setup SVM baseline algorithm outperforms other neural network models. Precision of the SVM model are significantly higher than the precision scores of any other machine learning model. In general, accuracies of LSTM models are better than CNN models according to the results of this experimental setup. Increasing the number of LSTM layers and number of LSTM unit cells in a layer didn't improve the accuracy. Even though, increasing the number of filters in a CNN layer and the number of CNN layers improve the classification accuracy. Increasing the number of filters over 50 leads over-fitting.

Performances of machine learning models that only employ Sentinel NDVI time series data are reported in Table 3.2. Machine learning models implemented based on Sentinel-2 NDVI data are more accurate than the models that are implemented based PlanetScope NDVI time series data. LSTM neural network which consists of single

LSTM layer and 75 LSTM unit cells reports the best results for all accuracy measurement scores. All CNN models show low accuracies compared to LSTM Models. SVM baseline model outperforms the accuracies of all CNN models, but not the aforesaid LSTM model accuracy measurements except the precision.

Accuracies of LSTM classification models built exploiting both PlanetScope NDVI data and Sentinel-2 LSWI data didn't outperform the accuracies of LSTM models that are built based on Sentinel-2 NDVI data (see Table 3.3). Accuracies of every CNN model is significantly higher compared to CNN models used in previous approaches. Maximum accuracy is reported by the CNN model that is included two CNN layers and 50 filters in each layer. Moreover, the best accuracy results of CNN have only a subtle difference with the best accuracy metric measurements of LSTM network. LSTM models with two layers reported better results over the LSTM models with one layer. On the other hand, LSTM models that are used in the previous two experiments conducted only using NDVI data show a completely opposite behavior. The performances of LSTM neural networks that consist of LSTM cell units less than 100, show notably lower results of less than 70%. Thus, we use 100, 125 and 175 as the number of LSTM cell units in hyper parameter settings. Two LSTM models and three CNN models surpass the accuracies of SVM model in terms of recall, F1-score and accuracy, but the SVM reports the best precision value among all ML models.

Neural networks implemented based on Sentinel-2 NDVI data and Sentinel-2 LSWI time series achieve the highest accuracies among all four experimental setups (see Table 3.4). Unlike in the previous experiments, CNN model that comprises with two layers and 50 filters in each layer, gained the highest scores nearly 98 percent for all accuracy measurements. However, LSTM model with two LSTM layers and 150 LSTM cell units achieve an accuracy which is very closed to the accuracy of aforesaid CNN model.

Overall, the neural network models implemented based on Sentinel-2 time series data outperforms the neural network models implemented based on PlanetScope time series datasets and combined time series datasets of PlanetScope and Sentinel-2. As

shown in Figure 3.2 and Figure 3.3, NDVI time series derived from PlanetScope are more complex than the time series derived from Sentinel-2 since the number of PlanetScope images used to derive PlanetScope NDVI time series is higher than the number of Sentinel-2 images used to create Sentinel-2 NDVI time series. As a result, it is challenging to achieve higher results by training machine learning models incorporating PlanetScope NDVI time series data with the similar hyper parameter settings used for the machine learning models implemented on Sentinel-2 NDVI data. Accuracies of LSTM neural networks are better compared to both 1D CNN model and SVM baseline model in most cases proving the potential of LSTM models for remote sensing time series analysis. Nevertheless, the capabilities 1D CNN in terms of time series classification have not been gained much attention in previous remote sensing research works, 1D-CNN models also show promising results with minor differences compared to LSTM models. Another significant observation is that the precision of SVM models in every experimental setup is nearly equal to the precisions of neural network models in which the results are the best in the respective experimental setup.

As showed in Figure 3.6, Figure 3.7 and Figure 3.8, variations of cultivated paddy extent estimation accuracies diverge from the variations of paddy classification accuracies. Even though the best accuracy of paddy land classification is reported by the model that used Sentinel-2 NDVI and Sentinel-2 LSWI time series. In contrast, the results of using only Sentinel-2 time series are poor in terms of estimating cultivated paddy extent. Models that employ only Sentinel-2 data fail to identify boundaries of paddy fields accurately. Pixels that cover the boundaries of paddy fields often represent a mixture of land types as the resolution of Sentinel-2 imagery is 10m. As mentioned above, we avoided picking boundary pixels for training ML models as it adds noises to the data set. But Sentinel-2 based estimations are affected by those noisy pixels in the test scene which do not represent paddy fields purely. It results in increase the number of false positives while the inability of classifying boundary pixels increases the number of false negatives. Consequently, balancing the effect of false positives and false negatives, the closest quantitative total cultivated paddy extent estimation is given by the model that used Sentinel-2 NDVI and

Sentinel-2 LSWI data. In contrast, model based on PlanetScope NDVI data and Sentinel-2 LSWI data, output approximately accurate paddy extent with a least impact from false positive and false negative


*Figure 3.6 Paddy extent estimation for a paddy field in dry zone. Boundaries of cultivated paddy areas are highlighted in (a). (b) shows the estimation given based on only PlanetScop NDVI data. Sentinel-2 NDVI data used in (c) while both PlanetScope NDVI and Sentinel-2 LSWI data are used in (d). Sentinel-2NDVI and Sentinel-2 LSWI data are used in (e)*

*Table 3.5 Paddy Extent Estimation Performance Evaluation*

| | Actual Extent | Planetscope NDVI | Sentinel-2 NDVI | Planetscope NDVI + Sentinel-2 LSWI | Sentinel-2 NDVI + LSWI |
|---|---|---|---|---|---|
| Region 1 (Dry Zone) | 25.47 | 26.9 | 27.33 | 26.61 | 27.14 |
| Region 2 (Intermediate Zone) | 37.62 | 30.61 | 21.01 | 40.09 | 36.61 |
| Region 3 (Wet Zone) | 8.96 | 11.81 | 11.54 | 8.71 | 5.29 |
| Total Extent | 72.05 | 68.32 | 69.88 | 75.41 | 69.04 |
| Error in Total Extent Est. | | -3.73 | -12.17 | 3.36 | -3.01 |
| Aggregated Region-wise Error | | 10.29 | 21.05 | 3.86 | 6.35 |


*Figure 3.7 Paddy extent estimation for a paddy field in intermediate zone.*



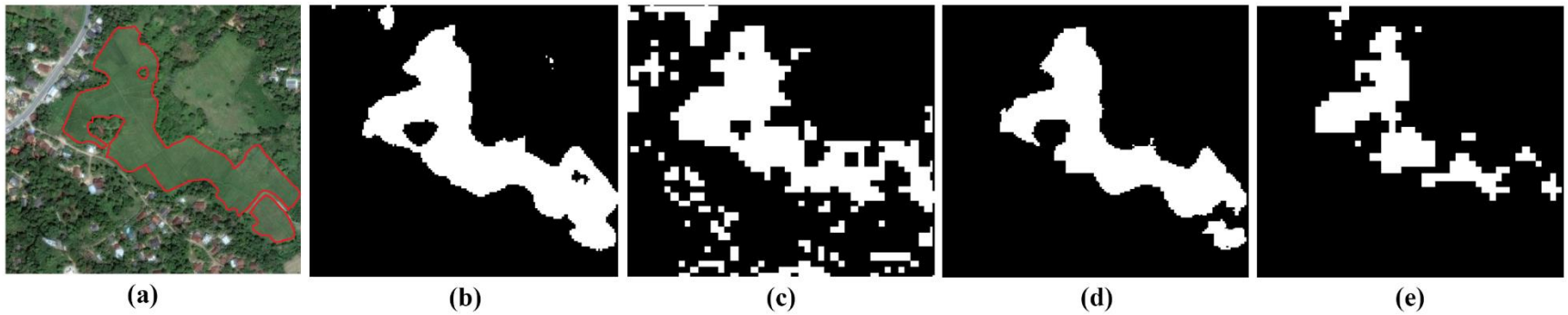


*Figure 3.8 Paddy extent estimation for a paddy field in wet zone.*

### 3.2 Cultivated paddy extent detection using active remote sensing data

As mentioned earlier, there are limitations of using traditional passive remote sensing data resources to examine and analyze cloud-prone agricultural areas, particularly during rainfall onsets. For instance, during the early phase of 2018-2019 Maha Season, most of the PlanetScope scenes that cover Matara District are hugely affected by cloud contamination. Thus, deriving proper time series including early growing season is not viable for those areas. Active remote sensing has been proved to be useful in above circumstances since radar beams penetrate through clouds. Thus, we implement a machine learning model which can be used to estimate cultivated paddy extent, by making use of Sentinel-1 Synthetic Aperture Radar (SAR) data. Then we discuss the accuracy measurements reported by the model and the potential of SAR imagery for estimating cultivated paddy extent in cloud-prone areas.

#### 3.2.1 Study area

As previously mentioned, three agro-climate zones can be identified in Sri Lanka. Sri Lanka is further divided into small agricultural divisions by Ministry of Agriculture in order to enable monitoring the cultivation and guiding farmers in more granular level. Ministry of Agriculture assigns agricultural instructors to engage in those tasks in each of aforesaid small agricultural divisions. As previously explained, cloud-free scenes are rare in Matara district which is located at Wet Zone. Hence, the sample ground data about cultivated paddy fields are collected from Thihagoda and Kamburupitiya areas in Matara district (See Figure 3.9) covering 210 acres with the assistance of agricultural instructors.

#### 3.2.2 Study data

The dataset used in this experiment consists of 5000 data points. The dataset represents the pixels that cover cultivate paddy fields, non-paddy fields regions and abandoned paddy fields. Sample data points are picked in a way that the adjacent pixels in a image don't get included in the data set in order to attenuate the effect of overfitting, by avoiding repeating the indistinguishable time series behaviors within the data set. Pixels that cover boundaries of the paddy fields are not included in the

data set in order to avoid representing mixed time series characteristics (Those pixels cover paddy field areas as well as non-paddy field areas).

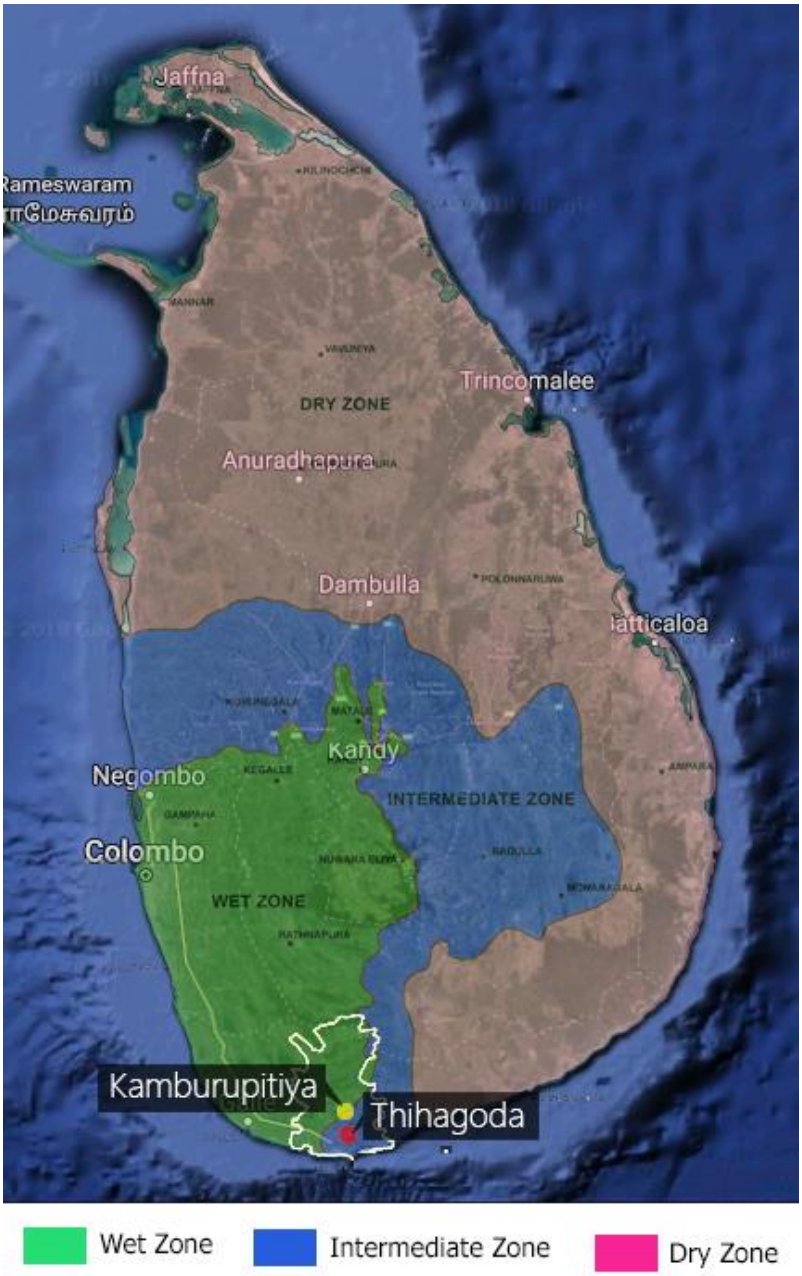


*Figure 3.9 Agro-ecological regions and locations of sample paddy fields. Samples are collected from Thihagoda and Kamburupitiya areas in Matara District (outlined in white)*

As previously stated, this experiment relies on Sentinel-1 Synthetic Aperture Radar data in order to mitigate the effect of cloud covers in Matara district. Sentinel-1 SAR imagery is obtained from Copernicus Open Access Hub of European Space Agency. Sentinel 1 satellites operate in four modes: Wide Swath (EW), Interferometric Wide

Swath (IW), Extra Wave (WV) and Stripmap (SM). We utilize data collected by Interferometric Wide Swath (IW) mode with dual polarization (VV+VH) scheme. Spatial resolution and temporal resolution of Sentinel-1 SAR imagery are 10m and 12 days respectively. Hence, time series characteristics between 10th of September 2018 and 15th of March 2019 is derived by using 10 SAR scenes.

### 3.2.3 Preprocessing steps

Sentinel-1 data are available in 3 categories according to level of processing:

- Level 0
- Level 1
- Level 2

We use Level-1 GRD product type for the experiment. Level-1 products are undergone various processing steps such as single look complex focusing and doppler centroid estimation etc. in order to convert Level 0 raw rata into usable data in various applications. Moreover, Speckles, a frequent issue with SAR imagery, are reduced in GRD products at the cost of reduced spatial resolution.

After obtaining Level 1 GRD products, a series of preprocessing steps is applied to every SAR image as shown in Figure 4.10. SNAP (Sentinel Application Platform) software is used to perform the preprocessing steps on SAR imagery. Firstly, orbit files are applied to enhance the geocoding accuracies of images [59]. The next step, Calibartion is essential to normalize the values of images, so that the images can be used for comparisons among time series. Third preprocessing step is applied to eliminate speckle noises in SAR images. Speckles inherently exist in SAR images due to interference among waves such as phase differences of waves, reflected from earth surface [60] [61]. The fourth step; terrain correction is required for SAR images since terrain variations (elevated and sloping terrains) affect accurate modeling of SAR scenes. Those terrain variations result geometric and brightness distortions in the scene [62]. In the subset step, we can define an area of interest to crop from each SAR scene. Feature extraction step is one of the most important preprocessing steps in any classification task. Texture analysis plays a vital role in feature extraction step of image classification applications. Gray level co-occurrence matrix (GLCM) can be

used to derive second order statistical texture features [63]. GLCM reflects the distributions of intensities comparing two pixels in an image located with a defined angle and a distance. Value of the element positioned at (i, j) in GLCM can be defined as the frequency of occurring two pixels, one of which the pixel gray-scale value is i, and the other gray-scale pixel value is j, and the angle and the distance of those two pixels theta and d (to indicate whether the two pixels located horizontally, vertically or diagonally) respectively [63]. Each of the element value of GLCM are calculated by applying the equation 12.

$$P_{(i,j)} = \frac{P_{(i,j,d,\theta)}}{\sum_{i=1}\sum_{j=1} P_{(i,j,d,\theta)}} \tag{12}$$

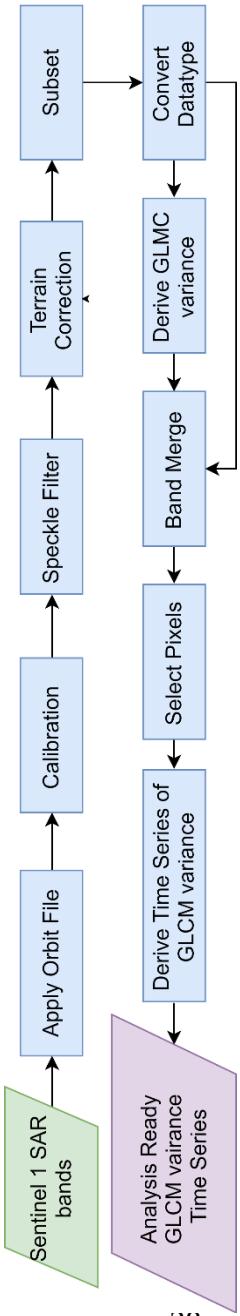


*Figure 3.10 Preprocessing steps applied for SAR images*

Statistical information about textures of the image can further analyze based on the derived GLCM. 14 kinds of textural features are introduced by Haralick et al. that can be obtained from GLCM (e.g.: entropy, correlation and contrast) [64]. In our experiment, we note that, 'sum variance' feature indicates a significant difference between the time series characteristics of cultivated paddy field areas and abandoned or non-paddy field areas. The equations for calculating the sum variance are showed in the equation 13 and equation 14 below. GLCM value band and all the spectral bands from the source image are merged into a single file in 'Band Merge' step. Then, the sample pixel set is selected for the study. GLCM sum variance values are then obtained for the selected pixels set. GLCM variance pixel values of 14 SAR images are put together and linearly interpolated to derive time series for each selected pixel in the final step of preprocessing procedure. Time series characteristics of a pixel that covers cultivated paddy field are shown in graph (a) in Figure 3.11. In contrast, graph (b) in Figure 3.11 shows time series characteristics of a pixel that covers non-cultivate paddy field.

$$f_{(1)} = -\sum_{i=2}^{2Ng} p_{(x+y)}(i)\log\{p_{(x+y)}(i)\} \tag{13}$$

$$Sum\ Variance = -\sum_{i=2}^{2Ng} \left(i - f_{(1)}\right)^2 p_{(x+y)}(i) \tag{14}$$

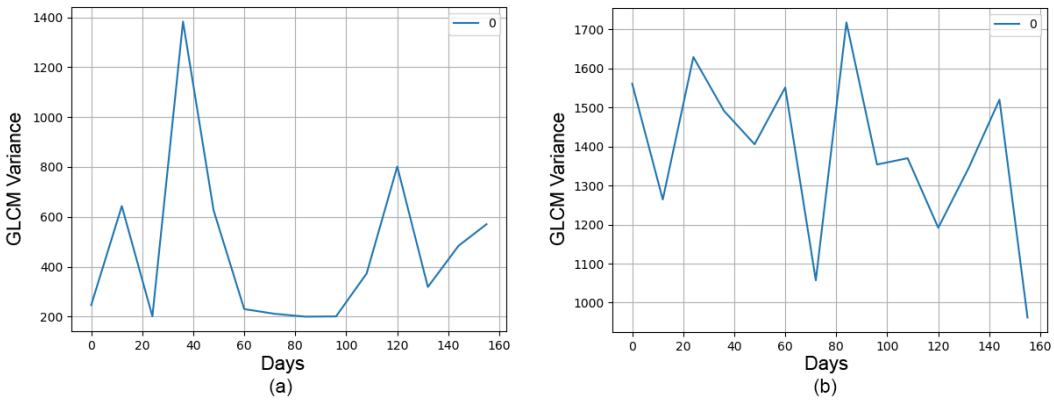


*Figure 3.11 GLCM Time Series graphs. Figure 3.11 (a) shows a GLCM Variance Time Series for a cultivated paddy pixel. Figure 3.11 (b) shows a GLCM Variance Time Series for a non-paddy field pixel*

### 3.2.4 Experiment setup

After completing preprocessing steps, we train a Convolution Neural Network with k-fold cross validation where k is 10 because the dataset is not large enough to divide into two datasets as training and testing. Since the entire dataset only consists of 5000 datapoints, the model is trained with 4500 datapoints and validated with 500 datapoints in each iteration. 1-dimension convolution neural network is used for the time series classification. The neural network model comprises with two convolution layers. Hyper parameter settings for the CNN layers are as follows:

- Number of filters: 20
- Kernel Size: 3

Maxpooling and Relu activation function layers are placed in between the convolution layers. The CNN model used in this experiment setup are implemented on Tensorflow platform using Keras library. We run the implemented CNN model on an unseen test dataset consists of pixels that belong to an entire sub-region in Matara District, in order to measure the potential of the model in terms of identifying cultivated paddy fields as shown in Figure 3.12.

### 3.2.5 Results

The best Convolution Neural network setup that is developed to identify cultivated paddy fields, shows accuracy of 96.20% with 90.88% of precision and 83.45% of recall. Paddy cultivation in Sri Lanka take places in small scattered paddy fields compared to larger continuous paddy fields in larger countries such as China. Utilizing Sentinel-1 SAR imagery for identifying cultivated paddy fields results low accuracy performances in the context of Sri Lanka. It hinders the ability to detect extremely small paddy fields which results subtle decrease in the recall (See Figure 3.12).

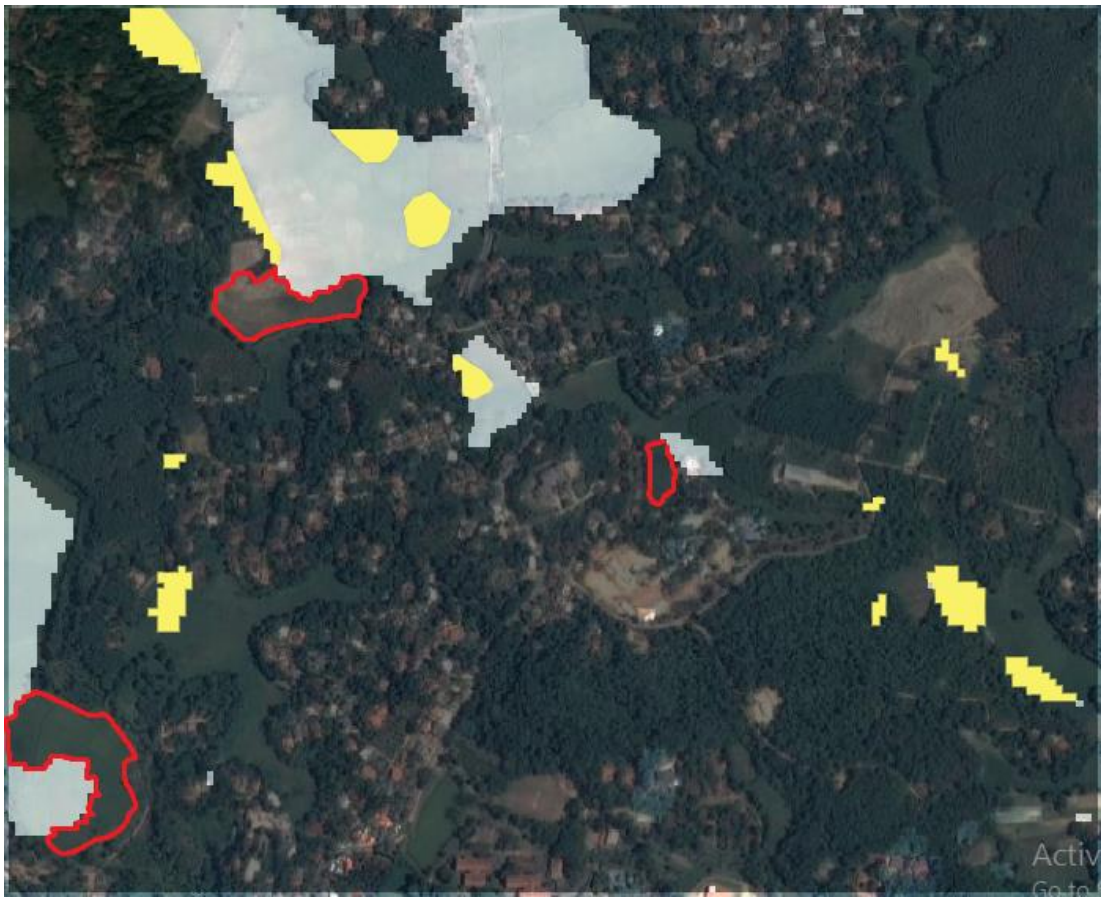

*Figure 3.12 Performance of CNN in terms of Cultivated Land Detection. Correctly identified cultivated paddy regions are colored in white. False positive pixels are colored in yellow. False negative regions are outlined in red.*

# 4. BROWN PLANTHOPPER ATTACKS DETECTION

In terms of pests and diseases detection, we mainly focus on detecting Brown Planthopper attacks in this study since it is the most devastating pest disease that causes yield losses [3]. Brown Planthopper (Nilaparvata lugens) has been recognized as one of the most catastrophic pest diseases in paddy cultivation around the world. The creature is a small brownish, sap-sucking insect that can fly over a large territory. The risk of escalating attacks at a blistering pace through a large region is high due to its ability of flying. The damages of BPH can be only visible when the plants suffer from hopperburn [65] in which the signs are similar to senescence [66]. But even low population of Brown Planthoppers can endanger the production by reducing tillering, crop vigor and increasing the percentage of unfilled grains [65].

It affects yield loses directly by feeding and indirectly by transmitting virus diseases and grassy stunt [67]. Approximately, 5% to 10% of rice production is lost annually due to BPH attacks in Sri Lanka [3]. Most of the farmers in Sri Lanka use resistive rice variants and pesticides to minimize the damage of Brown Planthoppers on paddy cultivation [68]. Nonetheless, severe BPH outbreaks are being reported in Asian countries including Sri Lanka in every year [69].

*Figure 4.1 Brown Planthopper*
*Source:http://www.knowledgebank.irri.org/training/fact-sheets/pest-management/insects/item/planthopper*

*Figure 4.2 Brown Planthoppers feed by sucking the juice in Paddy Plants*

*Figure 4.3 An area with hopperburn outlined in red*

We propose a novel approach for detecting brown planthopper attacks using PlanetScope data. We test the suitability of variety of indices derived leveraging the four spectral bands, for detecting BPH attacks. Then, we feed most suitable index parameters to a Support Vector Machine to detect BPH attacks in pixel level.

### 4.1 Study Area

Matara district that belongs to Wet Zone has been identified as one the highly threatened area by Brown Planthopper (BPH) attacks. Hence, the sample ground data regarding the areas affected by Brown Planthoppers are collected from Kamburupitiya and Thihagoda areas in Matara district with the aid of agricultural instructors (See Figure 4.9 above).

### 4.2 Study Data

The ground data is collected regarding the Brown Planthopper attacks occurred during 2018-2019 Maha Season. BPH attacks prevail as small scattered patches over the paddy fields. Therefore, only limited number of data points are available for this study of BPH attacks. We were able to pick 1000 data points for this experiment. Sample data points are chosen in a way that the nearby pixels in an image don't get included in the data set in order to minimize the effect of overfitting by avoid repeating the similar time series characteristics within the data set. Moreover, pixels that represents boundaries of BPH patches are not included in the data set in order to avoid representing mixed behavior of both healthy paddy plants and infested paddy plants.

We obtain 12 PlanetScope remote sensing scenes that were captured between 12$^{th}$ of December 2018 to 3$^{rd}$ of February 2019. Spatial resolution and temporal resolution of PlanetScope imagery are 3m and one day respectively. PlanetScope offers variety of product types according to the preprocessing steps that they have applied on and suitability for analytical applications. PlanetScope Analytic Ortho tile products are exploited in this experiment since those images are well-preprocessed for allowing researchers to use those products in analytical applications. Ortho tile product consists of four spectral bands:

- Blue (455-515) nm
- Red (590-670) nm
- Green (500-590nm)
- Near Infrared (780-860) nm

### 4.3 Preprocessing Steps

Changes of paddy plants due to BPH attacks, can’t be observed in SAR imagery since those changes do not affect the backscattering of radar beams. Hence, it is required to employ a passive remote sensing (optical) data resource to identify the changes in leaf structure due to Brown Planthopper attacks. PlanetScope remote sensing data resource is used as the passive remote sensing data resource in this experiment. Before releasing the PlanetScope Ortho tile products to general

analytical tasks, those data products go through a preprocessing pipeline as described earlier. Thus, every PlanetScope ortho tile product includes a file that indicates unusable pixel for analytic purposes due to cloud contamination and various other noises.

After acquiring PlanetScope imagery, every image is undergone through a preprocessing pipeline before passing it into analysis process (See Figure 4.1). First, we derive ratio indices and standard difference indices from the PlanetScope multi-spectral band file. Next, the data of four spectral bands; Red, Green, Blue and NIR are saved into separate from the files. Then we mark unusable pixels in all saved image files using the unusable data mask file in order to avoid considering low quality pixel values in the following analysis tasks. Each Ortho tiles image only covers 25 km by 25 km of small aerial footprint, so that to analyze some of the paddy fields, it is essential to merge several scenes (mostly two scenes) captured on same. Sample pixels are chosen and derived ratio indices, standard difference indices and spectral band time series for each of the selected pixels. Missing values in the time series of those raw spectral bands and indices are filled using linear interpolation. Unusable pixel values are eliminated to avoid representing false values in time series. As PlanetScope images are severely interfered by cloud contamination during early paddy growing season and Brown Planthopper attacks usually take place on December due to the favorable climate conditions, derived time series are limited to the latter half of the paddy cultivation cycle. Sample time series characteristics of the four spectral bands are showed in the Figure 4.2 while the sample time series characteristics of ratio indices are showed in the Figure 4.3. Figure 4.4 represents the sample time series characteristics of standard difference indices. The Ratio indices for a given day (d) between December 15th to Feb 15th are derived by applying the equation 15, 16 and 17 showed below.

$$(RI_{Green,p)})_d = \frac{(B_{Green})_d}{(B_p)_d} \qquad \text{where p = Blue, Red, NIR} \qquad (15)$$

$$(RI_{Blue,p)})_d = \frac{(B_{Blue})_d}{(B_p)_d} \qquad \text{where p = Red, NIR} \qquad (16)$$

$$(RI_{NIR,Red)})_d = \frac{(B_{NIR})_d}{(B_{Red})_d} \quad (17)$$

Standard difference indices for a given day (d) are calculated as equation 18, 19 and 20 showed below:

$$(SDI_{Green,p})_d = \frac{(B_{Green})_d - (B_p)_d}{(B_{Green})_d + (B_p)_d} \quad \text{where p = Blue, Red, NIR} \quad (18)$$

$$(SDI_{Blue,p})_d = \frac{(B_{Blue})_d - (B_p)_d}{(B_{Blue})_d + (B_p)_d} \quad \text{where p = Red, NIR} \quad (19)$$

$$(SDI_{NIR,Red})_d = \frac{(B_{NIR})_d - (B_{Red})_d}{(B_{NIR})_d + (B_{Red})_d} \quad (20)$$

SDI (NIR, Red) is also known as Normalized Difference Index (NDVI) that is often used in remote sensing research studies. Moreover, red, green, blue and NIR raw spectral bands are also taken into consideration for implementing the machine learning model. Then we subjected the derived indices and raw spectral bands to a linear correlation analysis in order to identify the most correlated spectral bands, ration indices and standard difference indices with Brown Planthopper. Parameters which show correlations greater than 0.7 are selected as inputs to the SVM model (See Table 5.1).

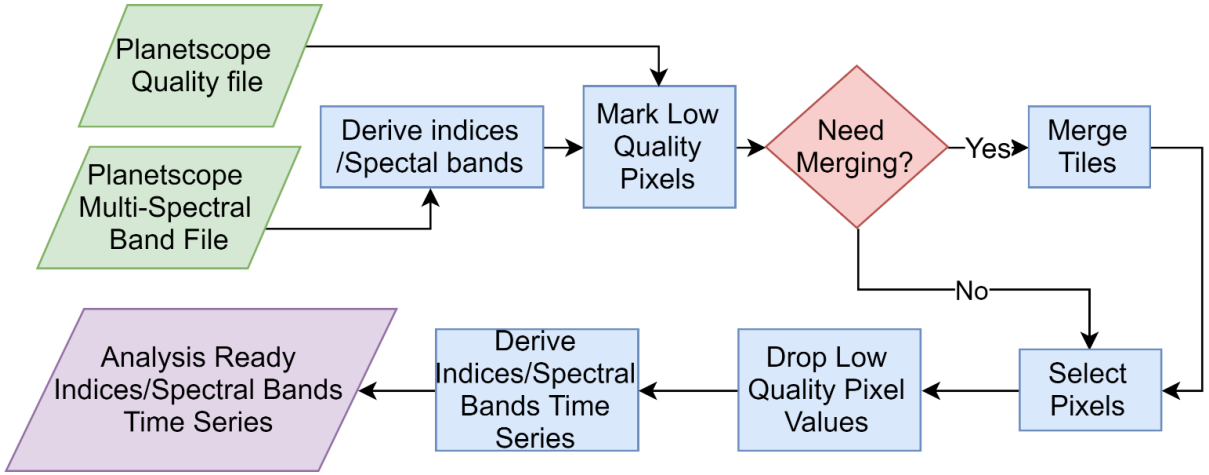


*Figure 4.4 Preprocessing Steps for BPH attack detection*



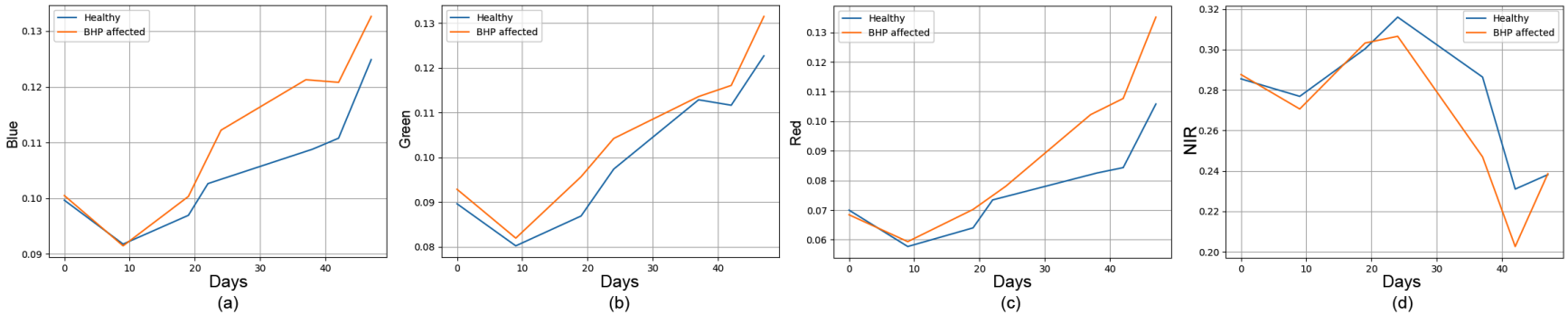


*Figure 4.5 Time series characteristics of spectral bands of a pixel covering a healthy paddy region are colored in blue while time series characteristics of spectral bands of a pixel covering BPH affected region are colored in orange. Graphs (a), (b), (c) and (d) represents Blue, Green, Red, and NIR bands respectively.*

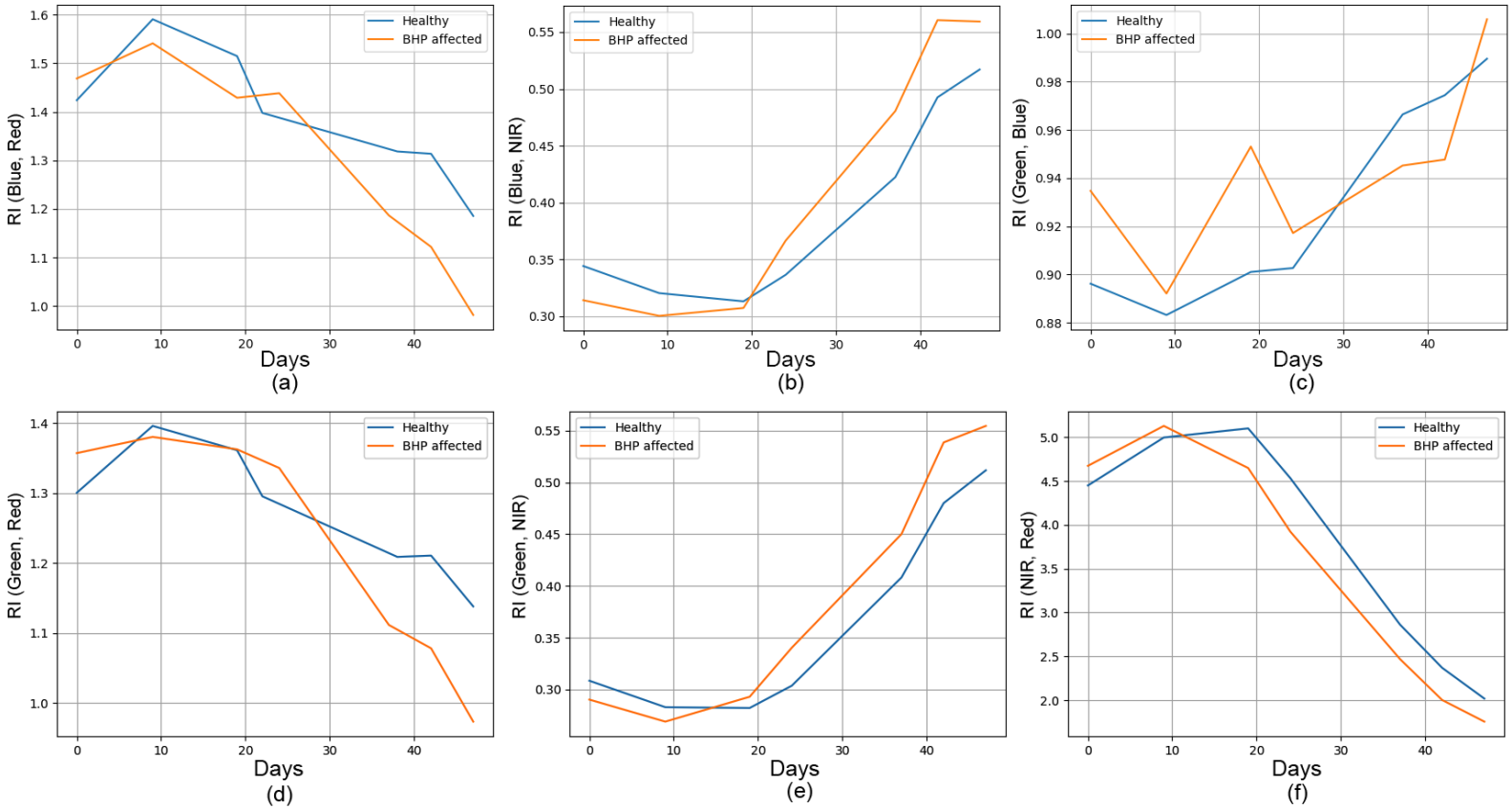


*Figure 4.6 Time series characteristics of ratio indices of a pixel covering a healthy paddy region are colored in blue while time series characteristics of ratio indices of a pixel covering BPH affected region are colored in orange. Graphs (a), (b), (c), (d), (e) and (f) represents RI (Blue, Red), RI (Blue, NIR), RI(Blue, Red), RI (Blue, NIR), RI (Green, Blue) RI (Green, Red), RI (Green, NIR) and RI (NIR, Red) indices respectively.*

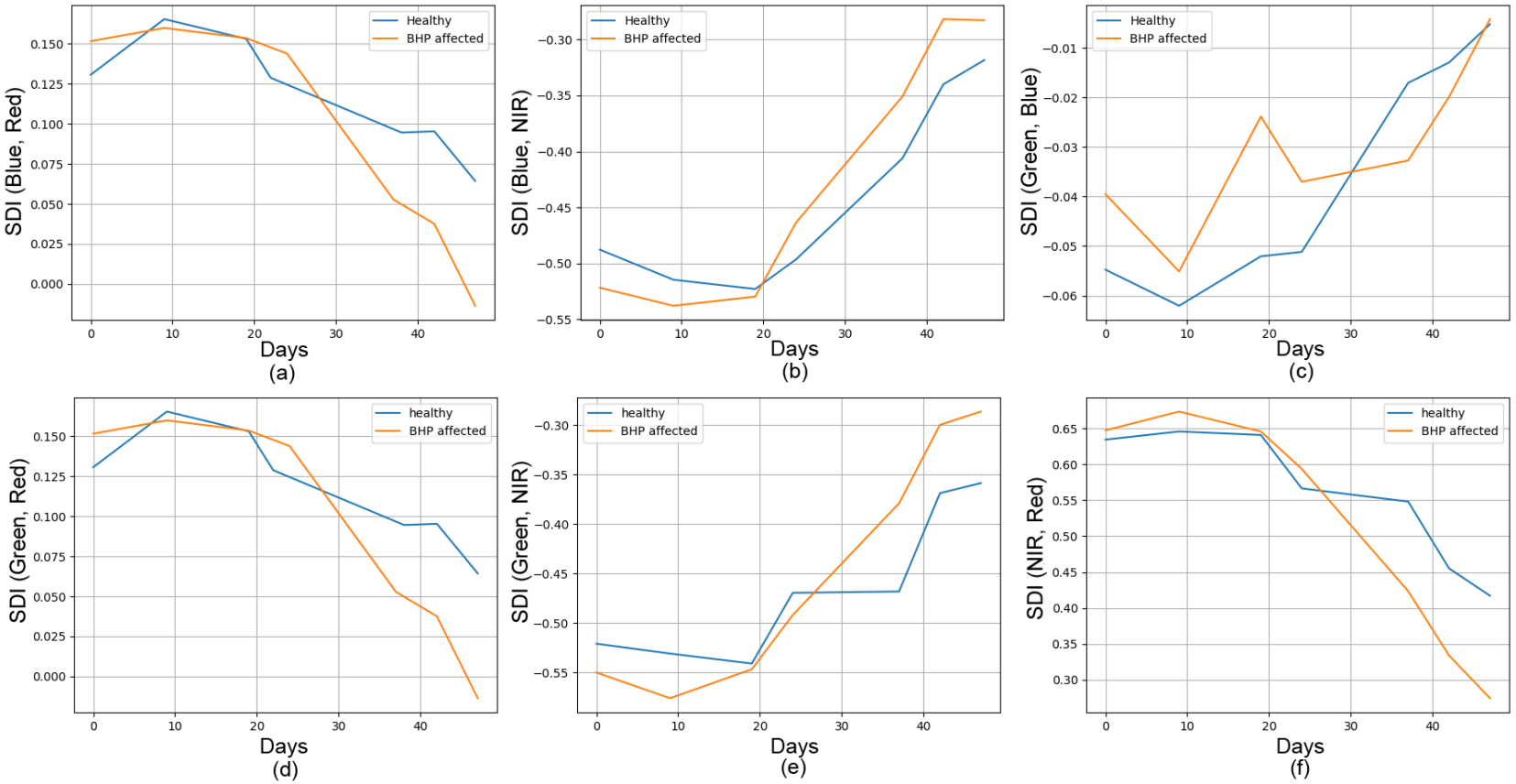


*Figure 4.7 Time series characteristics of standard difference indices of a pixel covering a healthy paddy region are colored in blue while time series characteristics of standard difference indices of a pixel covering BPH affected region are colored in orange. Graphs (a), (b), (c), (d), (e) and (f) represents SDI(Blue, Red), SDI(Blue,NIR), SDI(Green, Blue), SDI(Green, Red), SDI(Green, NIR) and SDI(NIR, Red) indices respectively.*

*Table 4.1 Index attributes with correlations greater than 0.7*

| Index | Selected Index | Index | Selected Days |
|---|---|---|---|
| RI (Blue, Red) | - | SDI (Blue, Red) | 31-36 |
| RI (Blue, NIR) | 39-47 | SDI (Blue, NIR) | 30-35 |
| RI (Green, Blue) | - | SDI (Green, Blue) | - |
| RI (Green, Red) | 32-36 | SDI (Green, Red) | 31-36 |
| RI (Green, NIR) | 39-48 | SDI (Green, NIR) | 39-48 |
| RI (NIR, red) | 40-47 | SDI (NIR, Red) | 38-48 |

Selected Days of RI (Blue, Nir) is mentioned as 39-47. It indicates that RI (Green, NIR) attribute of $39^{th}$ day to $47^{th}$ day has correlations greater than 0.7.

## 4.4 Experiment Setup

Only few PlanetScope pixels are often sufficient to cover a single Brown Planthopper patch. Therefore, the dataset only consists of 1000 datapoints. Thus, in the experiment, we used a Support Vector Machine model instead of Con. According to the correlation analysis done in the preprocessing procedure, 62 ration index and standard difference index parameters are selected to feed as inputs to the SVM model.

Hyper parameters of the Support Vector Machine are as follows:

- Gamma: 0.01
- Kernel: Radial Basis Function Kernel (RBF)

In general, 5% of paddy field areas are annually affected from Brown Planthopper attacks in Matara district. As in previous experiment setups, 10-fold cross validation is followed in order to split the dataset into training and testing dataset to mitigate the effect of the size of the dataset.

The main objective of this phase of the study is providing efficient and more precise alternative methodology for the government to detect the infested areas by BPH attacks. Hence, the implemented model is tested with an unseen test dataset covering a small area in Kamburupitiya region. Firstly, the non-cultivated paddy fields are filtered out by using the CNN model built using the Sentinel-1 SAR imagery. Then,

only the pixels that are identified as cultivated paddy areas are fed into the Support Vector Machine model for the classification. Alternatively, another test is carried out by feeding all the pixels in the selected area to SVM model without filtering the non-cultivated paddy fields pixels. Results of those two tests are depicted in the Figure 4.5 and Figure 4.6.

*Figure 4.8 Identified damaged regions by the SVM model. Area covering true positive results is outlined in red. False positive results are colored in yellow. False negative results are colored in blue.*

*Figure 4.9 Results of the SVM mdel after filtering out non-cultivated areas using the results of cultivated paddy fields dectection model.*

### 4.5 Results

Time series characteristic of raw spectral band reflectance, ratio indices and standard indices are showed in the Figure 4.2, 4.3 and 4.4 respectively. Adult Brown Planthoppers feed on rice plant by sucking the juice from the base of the rice plant. Feeding Brown Planthoppers shrink pigment content of leaves such as Chlorophyll and Carotenoid that are responsible for photosynthesis process in healthy plants [70]. This feeding process results alterations in reflectance of visible spectral regions. Photosynthetically active pigments absorb energy from red and blue spectral bands and reflect energy in green wavelengths [71]. After the rice plants get matured, growth of yellowish pinnacle may surpass the reflectance of green leaves of the rice plants. Thus, in contrast to N. R. Prasannakumar et al. discussed in one of his study [72], a higher reflectance of green spectral band in healthy plants may not be observed during the matured stage of paddy plants, from the view angle of satellites. But it is worth noting that the deviation of the reflectance behaviors between the infested and healthy paddy regions is very subtle in green band. Higher difference of the reflectance between infested and uninfested paddy areas can be observed in blue and red spectral bands compared to green band. The reflectance of near infrared spectral band is weak in infested plants. It is caused by decreasing the photosynthesis process due to the decay of the cell structure and low pigment concentration. Due to the severity of the infection, reflectance of raw spectral bands, values of standard difference and ration indices of infected plants drastically diverge from the values of corresponding parameters in healthy plants. We only fed parameters that have population correlation over 0.7 for Support Vector Machine model as inputs. The population correlation coefficients between the raw reflectance and BPH attacks are not significant enough in any day of the season. Red band on day 29 reports the highest correlation of 0.67. 0.60 on day 44 and 0.57 on day 23 are the highest correlation values corresponding to blue and green spectral bands. NIR band shows a negative coefficient and its highest correlation magnitude is 0.6 on day 43.

On the other hand, ratio indices and standard difference indices show more distinct discrimination between uninfested and infested paddy regions compared to raw spectral reflectance parameters. Substantial improvements of correlation coefficients

can be observed in standard difference indices and ratio difference indices except SDI (Green, Blue) and RI (Green, Blue). SDI (NIR, Red) is also known as the Normalized Vegetation Index (NDVI). It shows the highest negative correlation coefficient of -0.82. The reason for a such high correlation may be that NDVI is a good indicator of photosynthesis activities [73]. Further, a higher correlation coefficient can be observed in SDI (Blue, Red) and SDI (NIR, Red) compared to analogous ratio indices. The correlation values of standard difference indices during the period of study do not show any substantial enhancement or decline over corresponding ratio indices. RI (Green, NIR) reports the highest positive correlation coefficient of 0.81 among standard difference and ratio indices.

Support Vector Machine with above hyper parameter settings, reports 75.81% of accuracy, 94.28% of recall and 61.11% of precision. The slight drop of the recall is due to the mixed representation of the pixels that cover both paddy areas and non-paddy areas such as roads. As shown in Figure 4.5, the trained SVM model output some healthy paddy field areas and non-paddy field areas as damaged paddy fields resulting a significant drop of precision. Several reasons may affect to the low precision of the SVM model. Even though most of the parcels in a paddy field are harvested on a same date or a near dates, some parcels in a paddy field are harvested very early by farmers due to various reason. One of the main reasons behind such early harvesting is preventing the further prevalence of BPH attack into healthy paddy plants. Straws left in paddy field after harvesting, result similar spectral reflectance behavior as the paddy plants affected by BPH attacks.

# 5. AGRI AI PLATFORM

One of our objectives of this study is integrating our proposed systems with existing decision making procedure. Agri AI web platform is developed in order to fulfill that objective (See Figure 5.2). Agri AI system offers following functionalities for users:

- User can view the cultivated paddy fields in a map and the cultivated paddy extent estimation during a specific cultivation season and year.
- User can get estimations and view the cultivated paddy extent in a map for a specific area of the country. There are two ways of selecting specific area.
    - Uploading shape file specifying the area
    - Draw boundaries of the region in the map. (See Figure 5.3)
- User can get district wise cultivated paddy extent estimations and view cultivated paddy extent in a map**.**
- User view the damaged paddy field areas by BPH attacks in a map and the damaged paddy extent estimation during a specific cultivation season and year.
- User can get estimations and view the damaged paddy extent by BPH attacks in a map for a specific area of the country. There are two ways of selecting specific area.
    - Uploading shape file specifying the area
    - Draw boundaries of the region in the map
- User can get district wise cultivated paddy extent estimations and view cultivated paddy extent in a map**.**

Overall architecture of the Agri AI system is depicted in the Figure 5.1. Angular 6 JavaScript framework which was developed by Google, used for frontend development of the web application while the backend was implemented using Flask Python framework. MySQL is used as the database tool for the project. Geo spatial data manipulations are conducted using GDAL Python library.

The machine learning model for estimating the cultivated paddy extent is developed using Sentinel-1 SAR imagery in order to overcome the issue of cloud contamination. The trained models are saved so that, it can be used to detect

cultivated paddy areas for upcoming seasons without retraining machine learning models again. Classified areas are then saved in the disk as GeoTiff files. When a user requests estimation, boundaries of the selected region or the selected district name are sent to backend along with season details. Corresponding classified image is cropped using GDAL framework and calculated the cultivated paddy field extent or BPH attack prevalence. Finally, the estimations and the cropped image are sent to the frontend as the response.

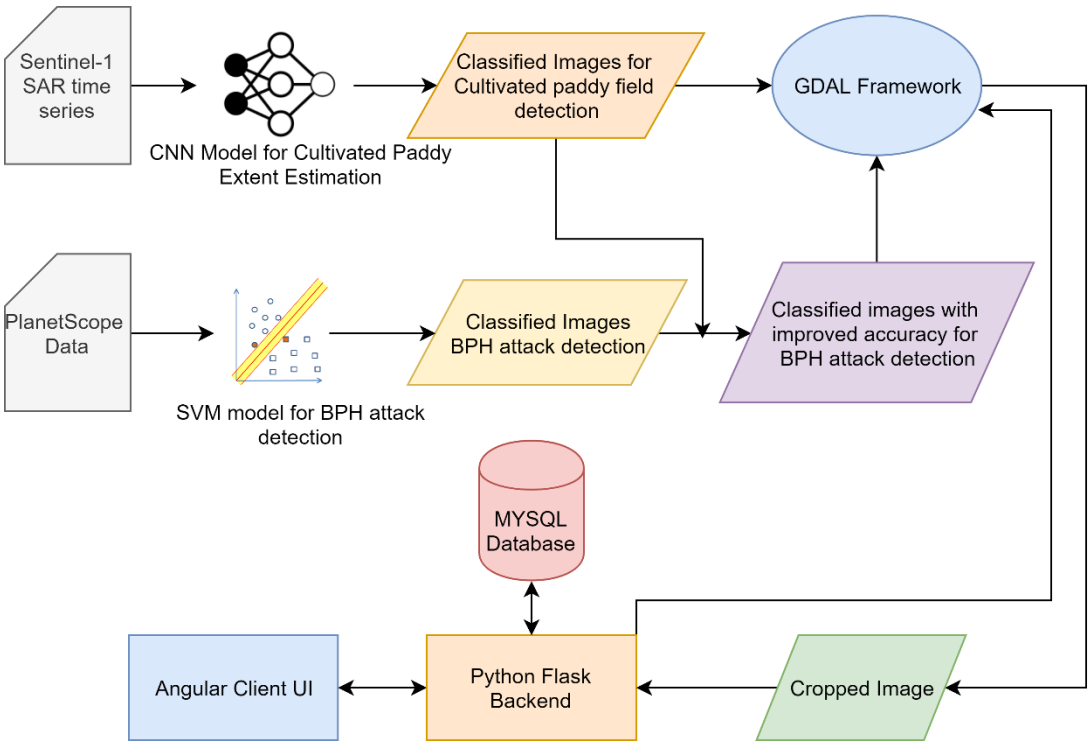


*Figure 5.1 Agro AI Architecture*

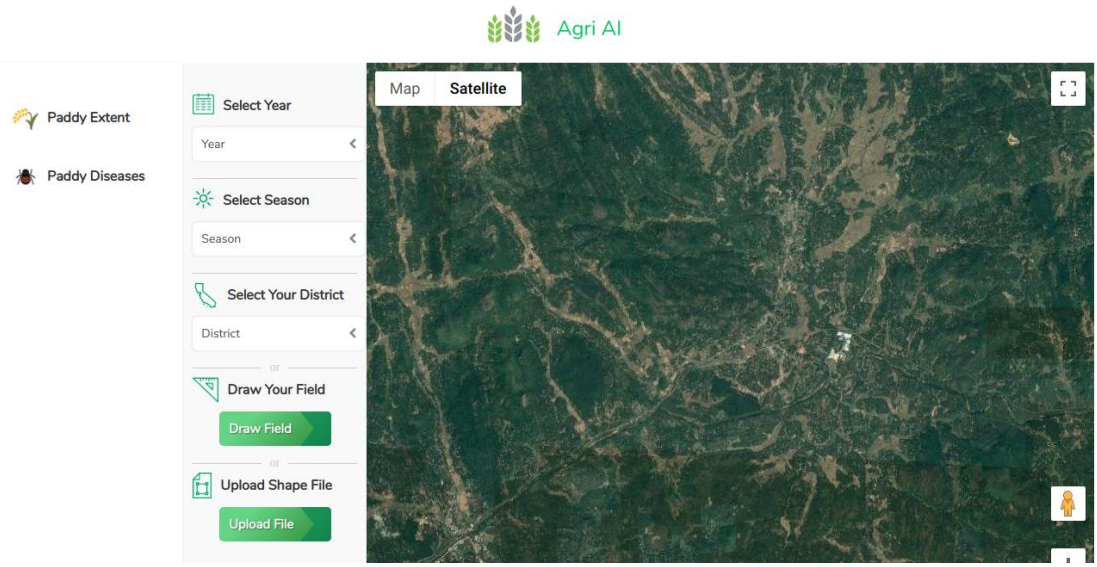


*Figure 5.2 Agri AI Web Interface*

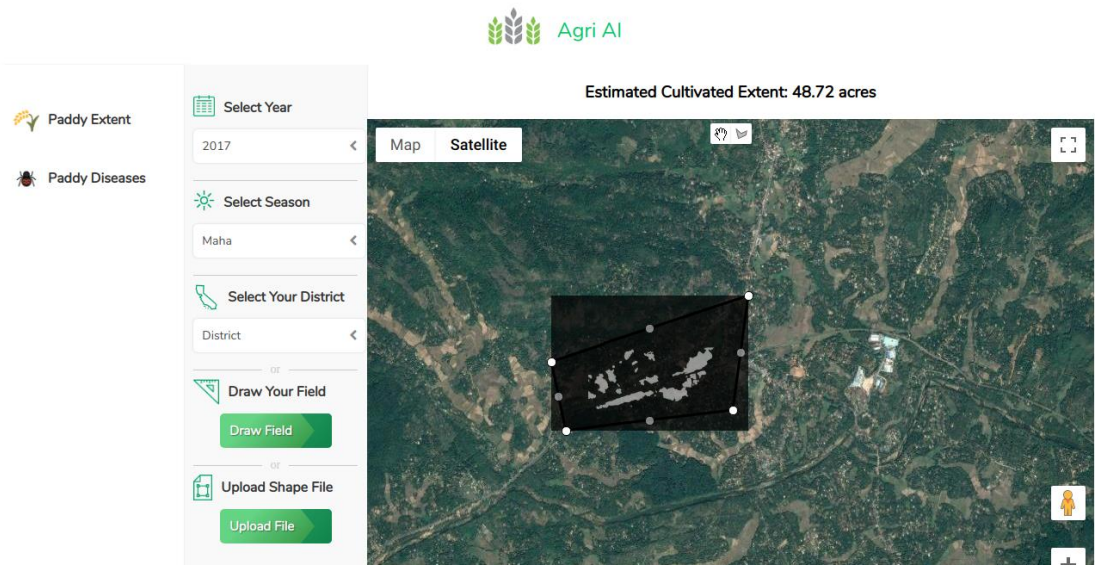


*Figure 5.3 User can draw a region in the map to get cultivated extent. Cultivated paddy regions are highlighted in white. Estimation for cultivated extent is displayed on the top.*

# 6. CONCLUSION

Both LSTM and 1D-CNN models can be used for classification tasks where it is necessary to take temporal characteristics into consideration. 1D-CNN can be considered as a good candidate for time series classifications even though LSTM is the state of the art neural network for time series classification tasks. If only the paddy pixel without mixed characteristics are used for training machine learning models, Sentinel-2 will report good paddy pixel classification accuracy despite the fact that resolution and revisit time are less than PlanetScope imagery. PlanetScope data can be suggested as a perfect optical remote sensing data resources for paddy land classification as well as paddy extent estimation in conditions where the cloud contamination is moderate or less. Most of the past studies only employed NDVI data for land classification applications. Results in this experiment shows a notable improvement of paddy land classification accuracies by incorporating LSWI data along with NDVI data. Even though the weather conditions of the three agro-ecological regions vary heavily, same machine learning models can be used without any change in any agro-ecological zone. It will greatly reduce the development time and the computation cost.

Heavy cloud contamination impacts the accuracies of the machine learning models that use passive remote sensing data, even though those models show a considerable robustness for mild and low cloud contaminations. In regions where heavy cloud contamination hits the accuracy of the paddy land classification, SAR imagery can be considered as a better alternative for classification tasks. Second experiment of the study shows promising results for cultivated paddy land mapping using SAR imagery time series. The suggested approach does not get affected by cloud contamination, so that it can be applied in anywhere without constraining on the weather conditions; particularly for the cloud-prone areas such as Wet Zone in Sri Lanka. Synthetic Aperture Radar imagery may be a better substitute for passive remote sensing data resources in crop mapping applications particularly in cloud-prone areas. Although SAR imagery are robust for cloud contamination, detecting very small paddy fields is nevertheless challenging for the approaches that use SAR imagery.

There was no significant study to measure the applicability of satellite based remote sensing resources to detect Brown Planthopper attacks. There were only few studies available to measure the severity level of BPH attacks based on statistical approaches and hyper-spectral remote sensing resources. Most of the previous research works for detecting various paddy diseases based on remote sensing were conducted only under laboratory conditions. In contrast, this study investigates the applicability of satellite based remote sensing for detecting BPH attacks. Overall, this study shows a great potential of applying optical remote sensing data with high temporal and moderate spatial resolution for BPH attack detection. Standard difference indices and ratio indices reports better correlations with BPH attacks over raw spectral reluctance. Support Vector Machine model identifies the BPH attacks with considerably better accuracy of 89%. But the field conditions such as early harvesting, hinder the precision of the model by a little. Precision of Support Vector Machine can be improved significantly by using the results of the previous experiment, masking out non-cultivated paddy field areas for the cost of slight drop of recall.

As previously mentioned, seasonal paddy yield is another vital parameter for precise decision making especially on exporting and importing decisions. Existing approaches of estimating crop yield and have errors and inefficiencies due to human involvement. The issues in the current decision making procedures could greatly be mitigated by extending this research study for estimating paddy yield. Authorities has defined average rice yield productions per hectare for different regions in the country. Hence, Phase 1 and Phase 2 calculation regarding the cultivated paddy extent and BPH attacks can be used to derive the paddy yield estimations in next steps.

Only four spectral bands are employed in the study to detect BPH attacks. Making use of more indices and spectral bands, particularly which indicate the water level (e.g.: Normalized Difference Water Index) of plant leaves may enhance the accuracy of the results. Furthermore, training the SVM model with a larger dataset will help with improving the accuracy.

Since the accuracy measures of the outputs of the combined approach of cultivated paddy area detection followed by BPH attack detection using SVM, are significantly high, proposed methodology can be suggested as a better substitute for traditional methodologies which involve human intervention for detecting the damaged areas by Brown Planthoppers. Moreover, the suggested combined approach of using SAR imagery and optical imagery to detect BPH attacks may have the capability of extending to detect other paddy diseases as well.

# REFERENCES


[1] M. Papademetriou, "RICE PRODUCTION IN THE ASIA-PACIFIC REGION: ISSUES AND PERSPECTIVES," [Online]. Available: http://www.fao.org/3/x6905e/x6905e04.htm. [Accessed 04 August 2019].

[2] Rice Research & Development Institute (RRDI), Bathalagoda, "Rice Cultivation," [Online]. Available: https://www.doa.gov.lk/rrdi/index.php?option=com_sppagebuilder&view=page&id=42&lang=en. [Accessed 04 August 2019].

[3] S. Madurangi, D. Ratnasekera, S. S. Senanayake, and P. Hemachandra, "Evaluation of brown planthopper Nilaparvata lugens (stal) resistance in Oryza nivara wild rice accessions found in Sri Lanka," in *Proceedings of International Forestry and Environment Symposium*, vol. 15, 2010.

[4] Department of Census and Statistics, "Rice Yield Estimations Sri Lanka," [Online]. Available: http://www.statistics.gov.lk/agriculture/Paddy%20Statistics/PaddyStats.htm. [Accessed 04 August 2019].

[1] W. G. Bastiaanssen, D. J. Molden, and I. W. Makin, "Remote sensing for irrigated agriculture: examples from research and possible applications," *Agricultural water management*, vol. 46, no. 2, pp. 137--155, 2000.

[2] C. Atzberger, "Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs," *Remote Sensing*, vol. 5, no. 2, pp. 949--981, 2013.

[3] S. Balaselvakumar and S. Saravanan, *Remote sensing techniques for agriculture survey*, 2006.

[4] M. S. Moran, Y. Inoue, and E. Barnes, "Opportunities and limitations for image-based remote sensing in precision crop management," *Remote sensing of Environment*, vol. 61, no. 3, pp. 319--346, 1997.

[5] M. Lewis, "Discriminating vegetation with hyperspectral imagery-what is possible?," 2001.

[6] N. Torbick, D. Chowdhury, W. Salas, and J. Qi, "Monitoring rice agriculture across myanmar using time series Sentinel-1 assisted by Landsat-8 and PALSAR-2," *Remote Sensing*, vol. 9, no. 2, p. 119, 2017.

[11] planet.com, "PLANET IMAGERY PRODUCT SPECIFICATION: PLANETSCOPE and RAPIDEYE," October 2016. [Online]. Available: https://www.planet.com/products/satellite-imagery/files/1610.06_Spec%20Sheet_Combined_Imagery_Product_Letter_E

NGv1.pdf. [Accessed 04 06 2019].

[12] A. Gatti, A. Galoppo, "Sentinel-2 Products," 14 March 2018. [Online]. Available: https://sentinel.esa.int/documents/247904/685211/Sentinel-2-Products-Specification-Document. [Accessed 04 06 2019].

[13] European Space Agency, "Sentinel-1 Operations," [Online]. Available: https://www.esa.int/Our_Activities/Operations/Sentinel-1_operations. [Accessed 04 06 2019].

[14] D. K. Bolton and M A. Friedl, "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics," *Agricultural and Forest Meteorology,* vol. 173, pp. 74--84, 2013.

[15] S. S. Panda, D. P. Ames, and S. Panigrahi, "Application of vegetation indices for agricultural crop yield prediction using neural network techniques," *Remote Sensing,* vol. 2, no. 3, pp. 673--696, 2010.

[16] B.D. Wardlow, S.L. Egbert, J.H. Kastens, "Analysis of time-series MODIS 250 m vegetation index data for crop classification in the US Central Great Plains," *Remote Sensing of Envrionment,* vol. 108, no. 3, pp. 290--310, 2007.

[17] B. D. Wardlow, S. L. Egbert, "Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the US Central Great Plains," *Remote sensing of environment,* vol. 112, no. 3, pp. 1096--1116, 2008.

[18] B. Zheng, S. W. Myint, P. S. Thenkabail, and R. M.Aggarwal, "A support vector machine to identify irrigated crop types using time-series Landsat NDVI data," *International Journal of Applied Earth Observation and Geoinformation,* vol. 34, pp. 103--112, 2015.

[19] Y. Shao, R. S. Lunetta, J. Ediriwickrema, and J. Iiames, "Mapping cropland and major crop types across the Great Lakes Basin using MODIS-NDVI data," *Photogrammetric Engineering & Remote Sensing,* vol. 76, no. 1, pp. 73--84, 2010.

[20] P. C. Doraiswamy, A. J. Stern, and B. Akhmedov, "Crop classification in the US Corn Belt using MODIS imagery," in *2007 IEEE International Geoscience and Remote Sensing Symposium*, 2007.

[21] R. Massey, T. T. Sankey, R. G. Congalton, K. Yadav, P. S. Thenkabail, M. Ozdogan, and A. J. S. Meador, "MODIS phenology-derived, multi-year distribution of conterminous US crop types," *Remote Sensing of Environment,* vol. 198, pp. 490--503, 2017.

[22] K. Chandrasekar, M. Sesha Sai, P. Roy, and R. Dwevedi, "Land Surface Water Index (LSWI) response to rainfall and NDVI using the MODIS Vegetation Index product," *International Journal of Remote Sensing,* vol. 31, no. 15, pp.

3987--4005, 2010.

[23] B.-C. Gao, "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space," *Remote sensing of environment,* vol. 58, no. 3, pp. 257--266, 1996.

[24] D. Chen, J. Huang, and T. J. Jackson, "Vegetation water content estimation for corn and soybeans using spectral indices derived from MODIS near-and short-wave infrared bands," *Remote Sensing of Environment,* vol. 98, no. 2-3, pp. 225--236, 2005.

[25] X. Xiao, S. Boles, J. Liu, D. Zhuang, S. Frolking,C. Li, W. Salas, and B. Moore III, "Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images," *Remote sensing of environment,* vol. 100, no. 1, pp. 95--113, 2006.

[26] L. Tingting and L. Chuang, "Study on extraction of crop information using time-series MODIS data in the Chao Phraya Basin of Thailand," *Advances in Space Research,* vol. 45, no. 6, pp. 775--784, 2010.

[27] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks,* vol. 61, pp. 85--117, 2015.

[28] M. Russwurm, M. Korner, "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[29] Z.C. Lipton, J Berkowitz, C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[30] Hu, F., Xia, G., Hu, J. and Zhang, L. "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing,* vol. 7, no. 11, pp. 14680--14707, 2015.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research,* vol. 15, no. 1, pp. 1929--1958, 2014.

[32] N. Kussul, M. Lavreniuk, S. Skakun, A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters,* vol. 14, no. 5, pp. 778--782, 2017.

[33] O. Abdeljaber, O. Avci, S. Kiranyaz, M. Gabbouj, D.J. Inman, "Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks," *Journal of Sound and Vibration,* vol. 388, pp. 154--170, 2017.

[34] B. Zheng, S.W. Myint, P.S. Thenkabail, R.M. Aggarwal, "A support vector machine to identify irrigated crop types using time-series Landsat NDVI data," *International Journal of Applied Earth Observation and Geoinformation,* vol. 34, pp. 103--112, 2015.

[35] F. Löw, U. Michel, S. Dech, C. Conrad, "Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using support vector machines," *ISPRS journal of photogrammetry and remote sensing,* vol. 85, pp. 102--119, 2013.

[36] B. Zheng, S.W. Myint, P.S. Thenkabail, R.M. Aggarwal, "A support vector machine to identify irrigated crop types using time-series Landsat NDVI data," *International Journal of Applied Earth Observation and Geoinformation,* vol. 34, pp. 103--112, 2015.

[37] Y. Shao Y, R.S. Lunetta, "Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points," *ISPRS Journal of Photogrammetry and Remote Sensing,* Vols. 78--87, p. 70, 2012.

[38] N. Kussul, M. Lavreniuk, S. Skakun, A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters,* vol. 14, no. 5, pp. 778--782, 2017.

[39] Z.Y. Liu, J.J. Shi, L.W. Zhang, J.F. Huang, "Discrimination of rice panicles by hyperspectral reflectance data based on principal component analysis and support vector classification," *Journal of Zhejiang University SCIENCE B,* vol. 11, no. 1, pp. 71--78, 2010.

[40] M. Bruke, D.B. Lobell, "Satellite-based assessment of yield variation and its determinants in smallholder African systems," *Proceedings of the National Academy of Sciences,* vol. 114, no. 9, pp. 2189--2194, 2017.

[41] N.A. Quarmby, J.R. Townshend, J.J. Settle, K.H. White, M. Milnes, T.L. Hindle, N. Silleos, "Linear mixture modelling applied to AVHRR data for crop area estimation," *International journal of remote sensing,* vol. 13, no. 3, pp. 415--425, 1992.

[42] F. Gao, J. Masek, M. Schwaller, F. Hall, "On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance," *IEEE Transactions on Geoscience and Remote sensing,* vol. 44, no. 8, pp. 2207--2218, 2006.

[43] T. Hilker, M. A. Wulder, N. C. Coops, J. Linke, G. McDermid, J. G. Masek, F. Gao, and J. C. White, "A new data fusion model for high spatial-and temporal-resolution mapping of forest disturbance based on Landsat and MODIS," *Remote Sensing of Environment,* vol. 113, no. 8, pp. 1613--1627, 2009.

[44] M. Immitzer, F. Vuolo, C. Atzberger "First experience with Sentinel-2 data for crop and tree species classifications in central Europe," *Remote Sensing,* vol. 8, no. 3, p. 116, 2016.

[45] X. Blaes, L. Vanhalle, P. Defourny, "Efficiency of crop identification based on optical and SAR image time series," *Remote sensing of environment,* vol. 96, no. 3-4, pp. 352--365, 2005.

[46] D.B. Michelson, B.M. Liljeberg, P. Pilesjö, "Comparison of algorithms for classifying Swedish landcover using Landsat TM and ERS-1 SAR data," *Remote Sensing of Environment,* vol. 71, no. 1, pp. 1--15, 2000.

[47] K. Jia, Q. Li, Y. Tian, B. Wu, F. Zhang, J. Meng, "Crop classification using multi-configuration SAR data in the North China Plain," *International Journal of Remote Sensing,* vol. 33, no. 1, pp. 170--183, 2012.

[48] Z. a. Z. M. Qin, "Detection of rice sheath blight for in-season disease management using multispectral remote sensing," *International Journal of Applied Earth Observation and Geoinformation,* vol. 7, no. 2, pp. 115--128, 2005.

[49] M. Zhang, Z. Qin, X. Liu, S.L. Ustin, "Detection of stress in tomatoes induced by late blight disease in California, USA, using hyperspectral remote sensing," *International Journal of Applied Earth Observation and Geoinformation,* vol. 4, no. 4, pp. 295--310, 2003.

[50] T. Kobayashi, E. Kanda, K. Kitada, K. Ishiguro, Y. Torigoe, "Detection of rice panicle blast with multispectral radiometer and the potential of using airborne multispectral scanners," *Phytopathology,* vol. 91, no. 3, pp. 316--323, 2001.

[51] G. A. Carter, "Responses of leaf spectral reflectance to plant stress," *American Journal of Botany,* vol. 80, no. 3, pp. 239--243, 1993.

[52] Z.Y. Liu, H.F. Wu, J.F. Huang, "Application of neural networks to discriminate fungal infection levels in rice panicles using hyperspectral reflectance and principal components analysis," *Computers and Electronics in Agriculture,* vol. 72, no. 2, pp. 99--106, 2010.

[53] Z. Yang, M. Rao, N. Elliott, S. Kindler, and T. Popham, "Differentiating stress induced by greenbugs and Russian wheat aphids in wheat using remote sensing," *Computers and Electronics in Agriculture,* vol. 67, no. 1-2, pp. 64--70, 2009.

[54] C.-M. Yang, C.-H. Cheng, and R.-K. Chen, "Changes in spectral characteristics of rice canopy infested with brown planthopper and leaffolder," *Crop science,* vol. 47, no. 1, pp. 329--335, 2007.

[55] N. Prasannakumar, S. Chander, and R. Sahoo, "Characterization of brown

planthopper damage on rice crops through hyperspectral remote sensing under field conditions," *Phytoparasitica*, vol. 42, no. 3, pp. 387--395, 2014.

[56] C.S. De Silva, E.K. Weatherhead, J.W. Knox, J.A. Rodriguez-Diaz, "Predicting the impacts of climate change—A case study of paddy irrigation water requirements in Sri Lanka," *Agricultural water management*, vol. 93, no. 1-2, pp. 19--29, 2007.

[57] F. Chollet, "Keras," [Online]. Available: http://keras.io/. [Accessed 25 07 2019].

[58] Google Brain Team, "Tensorflow," Google, 09 November 2015. [Online]. Available: https://www.tensorflow.org/. [Accessed 25 July 2019].

[59] A. McVittie, "SENTINEL-1 Flood mapping tutorial," SkyWatch Space Applications Inc., Feb 2019. [Online]. Available: http://step.esa.int/docs/tutorials/tutorial_s1floodmapping.pdf. [Accessed 22 05 2019].

[60] S.M. Ahmed, F.A Eldin, A.M. Tarek, "Speckle noise reduction in SAR images using adaptive morphological filter," in *2010 10th International Conference on Intelligent Systems Design and Applications*, 2010.

[61] J.-S. Lee, L. Jurkevich, P. Dewaele, P. Wambacq, and A. Oosterlinck, "Speckle filtering of synthetic aperture radar images: A review," *Remote sensing reviews*, vol. 8, no. 4, pp. 313--340, 1994.

[62] D. Small, "Flattening gamma: Radiometric terrain correction for SAR imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 8, pp. 3081--3093, 2011.

[63] L. Jia, Z. Zhou, and B. Li, "Study of SAR image texture feature extraction based on GLCM in Guizhou Karst mountainous region," in *2012 2nd International Conference on Remote Sensing, Environment and Transportation Engineering*, 2012.

[64] R. M. Haralick, K. Shanmugamet al., "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610--621, 1973.

[65] K. Sõgawa and M. Pathak, "Mechanisms of brown planthopper resistance in Mudgo variety of rice (Hemiptera: Delphacidae)," *Applied Entomology and Zoology*, vol. 5, no. 3, pp. 145--158, 1970.

[66] E. A. Backus, M. S. Serrano, and C. M. Ranger, "Mechanisms of hopperburn: an overview of insect taxonomy, behavior, and physiology," *Annu. Rev. Entomol.*, vol. 50, pp. 125--151, 2005.

[67] K. Sogawa, "The rice brown planthopper: feeding physiology and host plant interactions," *Annual review of entomology*, vol. 27, no. 1, pp. 49--73, 1982.

[68] C. Kudagamage and L. Nugaliyadde, "Present status and future direction of insect pest management in rice," *Rice congress Sri Lanka*, pp. 39--54, 1995.

[69] J. Xue, X. Zhou, C.-X. Zhang, L.-L. Yu, H.-W. Fan,Z. Wang, H.-J. Xu, Y. Xi, Z.-R. Zhu, W.-W. Zhouet al., "Genomes of the rice pest brown planthopper and its endosymbionts reveal complex complementary contributions for host adaptation," *Genome Biology*, vol. 15, no. 12, p. 521, 2014.

[70] Z.Y. Liu, J.J. Shi, L.W. Zhang, J.F. Huang, "Discrimination of rice panicles by hyperspectral reflectance data based on principal component analysis and support vector classification," *Journal of Zhejiang University SCIENCE B*, vol. 11, no. 1, pp. 71--78, 2010.

[71] C.-M. Yang, C.-H. Cheng, and R.-K. Chen, "Changes in spectral characteristics of rice canopy infested with brown planthopper and leaffolder," *Crop science*, vol. 47, no. 1, pp. 329--335, 2007.

[72] N. Prasannakumar, S. Chander, and R. Sahoo, "Characterization of brown planthopper damage on rice crops through hyperspectral remote sensing under field conditions," *Phytoparasitica*, vol. 42, no. 3, pp. 387--395, 2014.

[73] J. A. Gamon, C. B. Field, M. L. Goulden, K. L. Griffin,A. E. Hartley, G. Joel, J. Penuelas, and R. Valentini, "Relationships between NDVI, canopy structure, and photosynthesis in three Californian vegetation types," *Ecological Applications*, vol. 5, no. 1, pp. 28--41, 1995.