ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

COMPLETED RESEARCH:

CROPS CLASSIFICATION and YIELD ESTIMATION USING SPATIAL AND SPECTRAL FEATURES FROM REMOTE SENSING DATA

PI: TAGEL ABONEH
Co-PI: Wazih Mohammed
Addisu Mandfro

Adama, Ethiopia
4-Apri-22

ABSTRACT

Absence of advanced technological availability in the Ethiopia Agricultural sector. The estimation of average yield is done by the general crop estimation survey (GCES) technique which relies on the experiments designed at the time of harvesting period. In the current scenario, getting organized information about the crop status such as crop health, crop growth, acreages, and their estimated yields were difficult for many developing countries. In addition, the process of crop monitoring system was highly prone to error and results biased farming field survey outcome. In this regard, the processing and analyzing field survey data was a time-consuming process. To handle the aforementioned pitfalls, we proposed land use land-cover classification and yield estimation algorithms using a hyperspectral image data. To conduct the experiment, we have utilized Maximum likelihood, Random Forest and Support Vector machine learning algorithms. Three different Landsat image data have been acquired from the same area in a different time to understand the growth status of crops and their pattern in the specified area. A relevant spectral signatures has been extracted from the LandSat image data to conduct the experiment. From the experiment result, MLE, RF and SVM classified the land cover with the accuracy of 93, 98, and 97 respectively. From the experimental results, we concluded that the application of remote sensing technology in the agricultural sector has a significant contribution to improve production efficiency, managing and controlling crop growth, monitoring land use land cover, real-time data processing, analyzing and real-time decision making.

Keywords—Yield Prediction, Machine Learning, Remote Sensing, Crop Classification, Image Processing.

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

Introduction

Ethiopia is the second most populated countries in Africa next to Nigeria. About 85% of the population are participated in the agricultural domain and the sector have a significant contribution for national GDP [1]. The main challenge of the Ethiopian agricultural sector is lack of advanced technology and infrastructure, lack of automated system to collect and process real-time data, lack of well-organized information for the farmers. Another striking challenge was the processes of crop monitoring and resources management system were traditional and time-consuming [1] [2][3]. In addition, the average crop yield estimation procedure are mainly dependents on manually using the general crop estimation survey (GCES). This approach employed a randomly selected fields from the sampled villages to experiment the yield estimation for the total harvested output at national level. Domain expert argue that crop productions are heavily affected by environmental factors and volume of production also area dependent. Yield estimation [2] using randomly selected sampled data will result a hasty generalization due to the above-mentioned factors. Some of the major challenges are sample fields may not be representative, the procedure also takes too much time to process data, domain knowledge gaps to the interpretation problem and dynamic environmental changes affect to make a real-time decision. The advent of remote sensing technology around the 1970s, its great potential in the fields of agriculture have opened new opportunities to improve the decision-making process in the agricultural sector and contributes automate the existing manual system [3]. Remotely sensed hyperspectral [3] [4] image data has been used in the past for agricultural domain for different applications, among them the estimation of land cover with different crops is the major task. Besides the advancement of the technology in the domain area, the knowledge of spectral and spatial features about land-cover land-use provides a piece of valuable information to optimally utilize scarce natural resources and automated crop monitoring system. These challenges inherently motivate us to understand the spatial and spectral changes in crops and their impacts on the growth, land cover pattern analysis and yield estimation process. According to Bingfang Wu and Jihua [5] real-time information on agricultural production and yield estimation are essential to develop a strategic plan, to conduct market analysis, to formulate appropriate preventive mechanism, to normalize agricultural product import and export volume and to design a proper policy to support the sector.

Therefore, this research aimed to apply a machine learning algorithm experiment land cover and use analysis and yield estimation using multi-spectral image data. The main challenging was obtaining well label training dataset less than 10 meters spatial resolution to build machine learning model. Multi-spectral image provides rich information to characterize object on the surface of the earth by measuring the reflectance values [11]. Feature extraction, representation, and characterization of an object on the earth surface to classify the object has been done based on their spatial and spectral feature [6] [7]. The proposed model has been employed based on non-linear regression numerical algorithms to improve the prediction [10] accuracy.

Background

Satellite image has certain features like absorption, transmission, and reflectance [4] towards electromagnetic radiations. Most applications in remote sensing use these characteristics of the objects to process the data and obtain detail spectral and spatial information [1]. Recently, remote sensing become a hot research area in the agricultural to control and monitor crop health status and optimal resources management. But during the preliminary survey, we identified that Ethiopian agriculture need further research to automate the agriculture sector. In this regard, a limited numbers of researches outputs are reported in the domain area. Similarly, access to high quality Satellite data is a big challenge for researchers in the domain area. Due to the absence of the technology and irregular farming pattern, the sector was unable to address the food security [1] issues at national level. The current work is mainly focused to utilize machine learning methods [13] to perform two tasks: first to identify the land cover by classifying the type of crop on the selected study area and the second one is to perform yield estimation task. A review of related literature has been done and the different concerned organization were communicated to assess their requirements to implement the proposed machine learning models.

A proper yield estimation mechanism gives an important insight to understand the demand and supply variables and the gaps to be address to assure food security issues [5]. But the success of a computer-based crop classification system depends upon the careful selection of spectral features. To get insights, the research team has performed a literature review on documents, journals, periodicals; books and internet sites. According to [2], crop cutting experiments take long time to collect and analyze the health status with a higher level of

accuracy using the traditional technique. The traditional crop estimation techniques are well suited for small area estimation. The convention process is expensive and time-consuming processing. But, limitation of human visual sensor makes it difficult to crop disease at the early stage. The author suggests that better crop yield estimates can be produced by focusing on the property of spectral reflectance. To estimate crop yield following estimators were utilized by many researchers in the domain area:

- ✓ Crop yield estimator for the district without using remote sensing data.
- ✓ Crop yield estimator for the district based on post stratification using satellite data in the form of NDVI and RVI for stratification.
- ✓ A direct estimator of crop yield at grid level.

Problem statement

The conventional method of land-use land-cover analysis, crop monitoring and yield estimation were difficult and time-consuming process in the agriculture domain. In addition, the lack of availability of advanced technology such remote sensing image processing and limited access to well labeled remote sensing image data significantly affect the sector.

In addition, identifying suitable spectral features if the infra-red and near infra-red based images are used in computing the yield. Similarly, the spectral and spatial features need to be extracted from the image data to represent an object. Features extraction and representation were one of the challenging and time-consuming tasks in designing machine learning model.

Finally, after performing the classification task, the image data has been used for the purpose of yield estimation. Due to image spatial resolution problem, the current study apply open sources image data to compute the yield estimation task. Therefore, the current study have been proposed to address the following research questions:

Research Questions

The proposed land-use land-cover classification and yield estimation research was designed to address the following research questions:

1. What are the hyper-spectral features used to design a land-use land-cover classification model in the agricultural domain.

2. What are the robust techniques required for image pre-processing and Landsat image feature extraction purposes?

3. To evaluate the crop production loss rate against food security issues

4. What is the performance of the model for crop diseases detection and classification?

Objectives

To achieve the main purposes of the proposed system, the following specific activities has been accomplished:

- To obtain spectral images for the area under study during the season of Belg and/or meher on the differently cultivated croplands
- To identify the important spectral and spatial features of the land-cover under study and perform feature selection task
- To design a model for the purposes of land use land cover classifier using Satellite image data
- we have designed a mechanism to perform crop yield for each of the classified crops based on the vegetation index
- To compute the correlation of selected vegetation index with the ground observations

CHAPTER 2: LITERATURE REVIEW AND RELATED WORK

Introduction

Remote sensing can be defined as collection and interpretation of information about an object, area or event without any physical contact with the object. Aircraft and satellites are the common platforms for remote sensing of earth and its natural resources [6]. Thus, the goal of hyperspectral imaging is to obtain the spectrum for each pixel in the image of a scene, with the purpose of finding objects, identifying materials, or detecting processes.

According to Praveen, P. and his colleague [7] explained that Hyperspectral sensor [8] images allow users to observe how the surface of the Earth is altering swiftly, at local, regional, national, even global scales. Due to the detailed spectral information available from the hundreds of (narrow) bands collected by hyperspectral sensors, accurate discrimination of different materials is possible. This fact makes hyperspectral data a valuable source of information to be fed to advanced classifiers [9]. On the other hand, Jinru Xue and Baofeng Su [10] described that, remote sensed information can provide extremely useful insights for applications in environmental monitoring, biodiversity conservation, agriculture, forestry, urban green infrastructures, and other related fields.

According to Tao, R. hyper-spectral imagery (HSI) comprises hundreds of contiguous spectral bands, which enables to distinguish different objects with subtle spectral difference [11]. On the other hand, author argue that yield estimation can be further improve by employing remote sensing-based estimation techniques [12]. The author suggests that, better crop yield estimates can be produced by utilizing a robust spectral and spatial feature extraction techniques. Tao, R. and his colleague proposed fractional Fourier entropy (FrFE)-based hyperspectral anomaly detection method to extract feature from hyper-spectral image data [11].

Different organization have been communicated to assess a requirement in order to propose efficient machine learning models that can handle the existing pitfall. From the interviews of domain experts and collected data, we found that the classification of crops based on the spectral and spatial features is a near real-time and continuous process, which has important role in planning, monitoring growth status of crops, damage prevention and designing strategy for different secondary tasks in the sector. But, due to lack of high-resolution image data, we enforced to perform the land cover and land use analysis instead of doing classification

at crop levels. Remotely sensed image data contains rich information about absorption, transmission and reflectance value of each object [13] [14].

According Melgani, F. [12], relatively small number of acquisition channels that characterizes multi-spectral sensors may be sufficient to discriminate among different land-cover classes (e.g., forestry, water, crops, urban areas, etc.). However, their discrimination capability is very limited when different types (or conditions) of the same species (e.g., different types of forest) are to be recognized.

Hyperspectral data cover a wide spectral range from the visible to the short-wave infrared, resulting in hundreds of data channels. Thanks to this volume of information, it is feasible to deal with applications that require a precise discrimination in the spectral domain. Nevertheless, a large number of features can become a curse in terms of classification accuracy if enough training samples are not available, i.e. the Hughes phenomenon. The use of conventional statistical methods may not be adequate for classifying high-dimensional data and therefore more sophisticated classifiers need to be considered [13].

Land-use Land-Cover Analysis

Xue, L. and his colleague [14] inspired remote sensing technology for Land cover classification [15] using remotely sensed data is capable of generating information that can play an important role in forest resource inventory, agricultural monitoring [2], and environmental change. According to Xue, the use of complex machine learning algorithms in land cover remote sensing remains in its infancy and choosing the most suitable method for large-scale land cover mapping remains a challenge. A suite of machine learning classifier are available for image analysis such as Naive Bayes, Support Vector Machine (SVM) [12][13], Classification and Regression Trees (CART), Random Forest (RF), Gradient Boosting Machine (GBM), Neural Network (NN), and others. They have become increasingly popular within the field of remote sensing in recent years due to their applicability across large datasets and their ability to generate more accurate and consistent results. Generally, information contained in hyperspectral data allows the characterization, identification, and classification of the land-covers with improved accuracy and robustness. However, several critical problems should be considered in classification of hyperspectral data, among which: 1) the high number of spectral channels; 2)

the spatial variability of the spectral signature; 3) the high cost of true sample labeling; and 4) the quality of data [16].

Application of Remote Sensing image processing

Hyperspectral imaging has become a valuable remote sensing tool due to the development of advanced remote acquisition systems with high spatial and spectral resolution, and the continuous developments on more efficient computing resources to handle the high volume of data. For this reason, hyperspectral image analysis has found important uses in precision agriculture, where the health status [17][18] of crops [19] in various stages of the production process can be assessed from their spectral signatures [20] [5].

According to Shahid Mohammad, S. and his colleague [21] mentioned that, Machine learning whenever implemented have resulted in better outputs in terms of various metrics, sometime it may even exceed human expertise. Machine learning techniques can be categorized on the bases of learning mechanism. Machine learning algorithms due to their outstanding predictive power have become a key tool for modern hyperspectral image analysis. Therefore, a solid understanding of machine learning techniques has become essential for remote sensing researchers and practitioners.

In addition, Landsat image processing has used to maximize the productivity of this activity and minimize economic and food crop losses, various precision agriculture techniques to optimize yields [4] by managing production inputs and monitoring [3] plant health have been developed. Such applications include land-cover classification, cultivar identification, nitrogen level assessment, chlorophyll content estimation and the identification of various factors, such as the presence of pests, weeds, disease or pollutants [5]. Due to the advantages gained from sensing a large number of narrow spectral bands and a wider range of the electromagnetic spectrum. Moreover, HSI has been exploited in increasing number of applications [1] from remote and proximal sensing, chemical processes, medical imaging, and industrial processes to agricultural and environmental monitoring [22]. It is worth noting that hyperspectral images can be acquired using four different configurations single shot, area, line, and point scanning. Remote sensing technology has proved to be an efficient tool for monitoring crop growth and identifying crop diseases over the last several decades [23].

Similarly, Vegetation information from remote sensed images is mainly interpreted by differences and changes of the green leaves from plants and canopy spectral characteristics. The most common validation process is through direct or indirect correlations between VIs obtained and the vegetation characteristics of interest measured in situ, such as vegetation cover, LAI, biomass, growth, and vigor assessment. More established methods are used to assess VIs [23] using direct and geo-referenced methods by monitoring sentinel plants to be compared with VIs obtained from the same plants for calibration purposes [24][25][10]. Modern machine learning based vegetation parameter estimation approaches automatically learn the relationship between reflectance spectra and vegetation parameter of interest from training data.

Challenges in the domain area

When dimensionality (the number of bands) increases, with a constant number of training samples, a higher dimensional set of statistics must be estimated. In other words, although higher spectral dimensions increase the separability of the classes, the accuracy of the statistical estimation decreases. This leads to a decrease in classification accuracies beyond a number of bands. For the purpose of classification, these problems are related to the curse of dimensionality [9]. In addition, Classification accuracies can be highly influenced by the spatial resolution of the hyperspectral data. A higher spatial resolution can significantly reduce the mixed-pixel problem and detect more details of the scene. In addition, sufficient training samples to simplify the complexity of classification and prediction processes were another challenge in the domain of multi-spectral image processing [17].

CHAPTER 3: RESEARCH METHODOLOGY

METHODOLOGY

Multi-spectral image data or Satellite data gives us rich information to identify object from the surface of the earth. But Satellite data acquisition is a difficult task due in the case of Ethiopia. In this research study, we have employed an experimental research approach to conduct the current study. The following architecture describes the conceptual architecture of the proposed system.

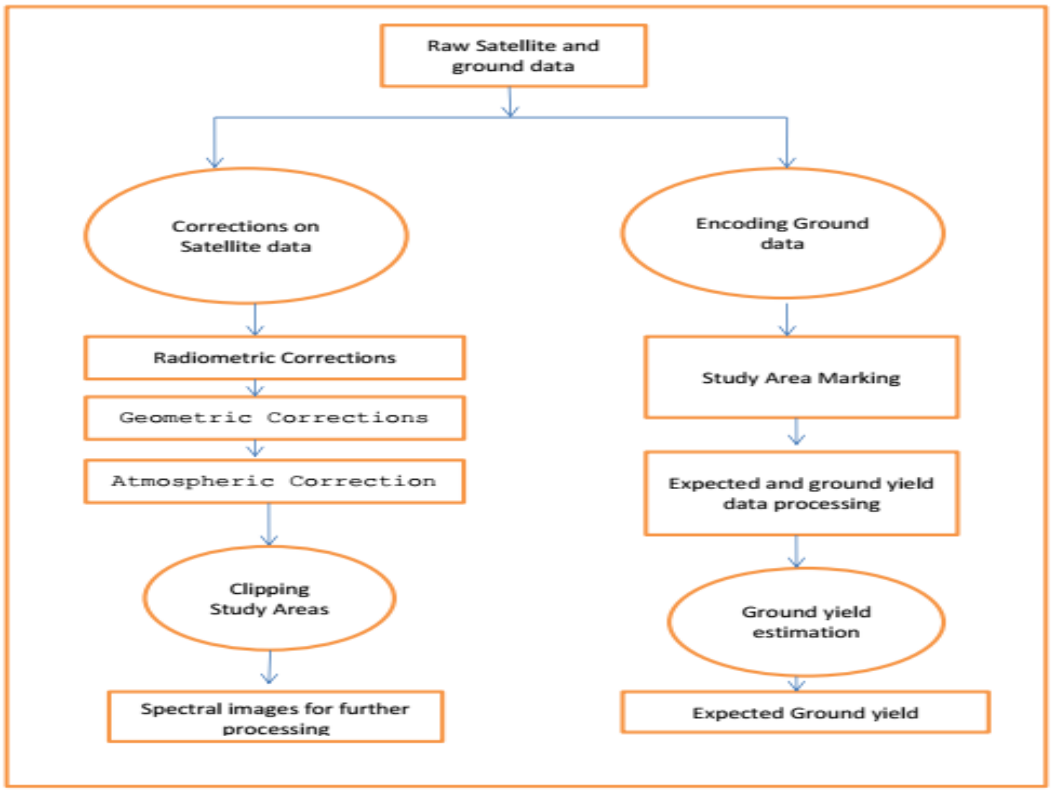

Figure 1:The general framework of the proposed system

Data collection

A survey has been conducted at Bishoftu agricultural research institutes to specify the data selection area. Based on the support and recommendation of domain experts, Yerer Sillasie area has been selected to collect Landsat image data. We have selected October to December 2018 for three months to acquire the Landsat image. In this period, the percentage of cloud cover becomes very minimal. On the other hand, the site is selected due to the availability of diverse land-cover.

Figure 2 below illustrated the geometric coordination of selected study area (Yerer Selassie) near Bishoftu Town Oromia Region, Ethiopia.

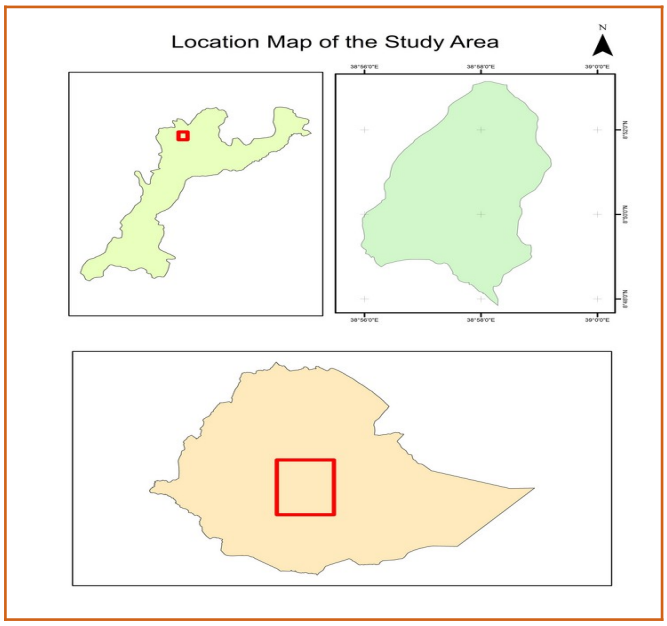

Figure 2:The selected study area

From the above figure the bottom most show Ethiopia, the upper left figure indicates Bishoftu Oromia and the upper right figure is the selected study area (Yerer Sillasie).

Landsat Data Sources

Yarer Selassie is located in Bishoftu east Shewa zone Oromia region of Ethiopia at geometric coordinate at 38°57’40.863"E and 8°50'55.016"N. Yerer Selassie bordered on the south by Dugda Bora, on the west by the West Shewa Zone, on the northwest by Akaki, on the northeast by Gimbichu, and on the east by Lome. Altitudes in this Woreda range from 1500 to over 2000 meters above sea level. Although the highest point in Ada'a, and in Misraq Shewa, is Mount Yarer which lies on the border with Akaki, Mount Zuqualla is also a prominent peak as well as a notable landmark, as the monastery of Saint Gebre Manfas Qeddus is located on it.

In order to conducted the experiment, we have utilized three image datasets collected in different time-frame (October 2018, November 2018 and December 2018) from the LANDSAT 8. All satellite images produced by NASA are published by NASA Earth Observatory and are freely available to the public. Satellite images are useful in meteorology, oceanography, fishing,

agriculture, etc. Although, there are different imaging satellite's we acquire our image from Landsat satellite. For the purpose of this study, we have used Band 2, Band3, Band4 and Band 5 spectrum to extract the relevant features used to build machine learning model. Band selection are mainly correlated with the problem domain and the purposes of the current study.

Table 1:Selected bands of Landsat8

Landsat OLI-TIRS bands	
30 m Coastal (0.43 – 0.45)	Band 1
30 m Blue (0.45 – 0.51)	Band 2
30 m Green (0.53 – 0.59)	Band 3
30 m Red (0.63 – 0.67)	Band 4
30 m NIR (0.85 – 0.88)	Band 5
30 m SWIR-1 (1.57 – 1.65)	Band 6
100 m TIR-1 (10.60 – 11.19)	Band 10
100m TIR-2 (11.50 – 12.51)	Band 11
30m SWIR-2 (2.11 – 2.29)	Band 7
15m Pan (0.50 – 0.68)	Band 8
30m Cirrus (1.36 – 1.38)	Band 9

Data and data source

To analyze the patterns of land-cover land-use on the study area, we have collected Landsat image data during above-mentioned time frame. The satellite image data in each period have its own data ID and unique descriptions.

Table 2:Hyperspectral Image data source

No	Data source	Data ID	Acquisition date	Path, Row	Size on disk
1	Landsat-8 OLI/TIRS	LC81680542018285LGN00	2018-10-12	168,54	1.3GB
2		LC81680542018317LGN00	2018-11-13	168,54	1GB
3		LC81680542018349LGN00	2018-12-15	168,54	.987GB

Multi-spectral image processing

In case of multi-spectral image processing, the raw image data cannot be used for classification or prediction purpose. The data pre-processing techniques has been used to extract the spectral and spatial features from satellite images. The spatial data pre-processing has been done using conversion of digital number in to Reflectance and Topographic Correction. The pre-processing and post processing were carried out to enhance the quality of the images. In this study, we have used the ERDAS Imagine and QGIS tools to extract the spectral feature and visualization purpose. The main satellite image pre-processing tasks has been discussed as follows:

OLI Top of Atmosphere Reflectance:

To estimate from the Landsat-8 OLI/TIRS TIR band, DN of sensors were converted to spectral radiance. The 16-bit integer values in the L1 product can also be converted to TOA reflectance. The following equation were used to convert Level 1 DN values to TOA reflectance:

$\rho\lambda' = M\rho * Qcal + A\rho$, where:

$\rho\lambda'$ = TOA Planetary Spectral Reflectance, without correction for solar angle (unit less)

$M\rho$ = Reflectance multiplicative scaling factor for the band

(REFLECTANCEW_MULT_BAND_n from the metadata).

Qcal = L1 pixel value in DN

$A\rho$ = Reflectance additive scaling factor for the band (REFLECTANCE_ADD_BAND_N from the metadata).

Note that $\rho\lambda''$ is not true TOA Reflectance, because it does not contain a correction for the solar elevation angle.

where: $\rho\lambda$ = Reflectance, $\rho'\lambda$ = TOA Planetary Reflectance (Unitless)

$\sin(\theta)$ = Solar Elevation Angle (from the metadata, or calculated)

Before performing the classification of the RS data, it is important to pre-process the data to correct the error during scanning, transmission and recording. It refers to the functions which are frequently performed to improve geometric and radiometric qualities of the images. Typically, the pre-processing steps are:

Radiometric correction and Topographic Correction

Atmosphere correction is used to modify DN (digital number) values to account for noise i.e. contributions to DN values. Radiometric correction is used avoid radiometric error and distortions. When the emitted or reflected electromagnetic energy is observed by a sensor on board an aircraft or spacecraft, the observed energy does not coincide with the energy emitted or reflected from the same object observed from a short distance. This is due to the sun's azimuth and elevation, atmospheric condition such as fog or aerosols, sensor's response etc. which influence the observed energy. Therefore, in order to obtain the real radiance or reflectance, those radiometric distortions must be corrected. The following architecture depicted the procedure of data correction to the required output.

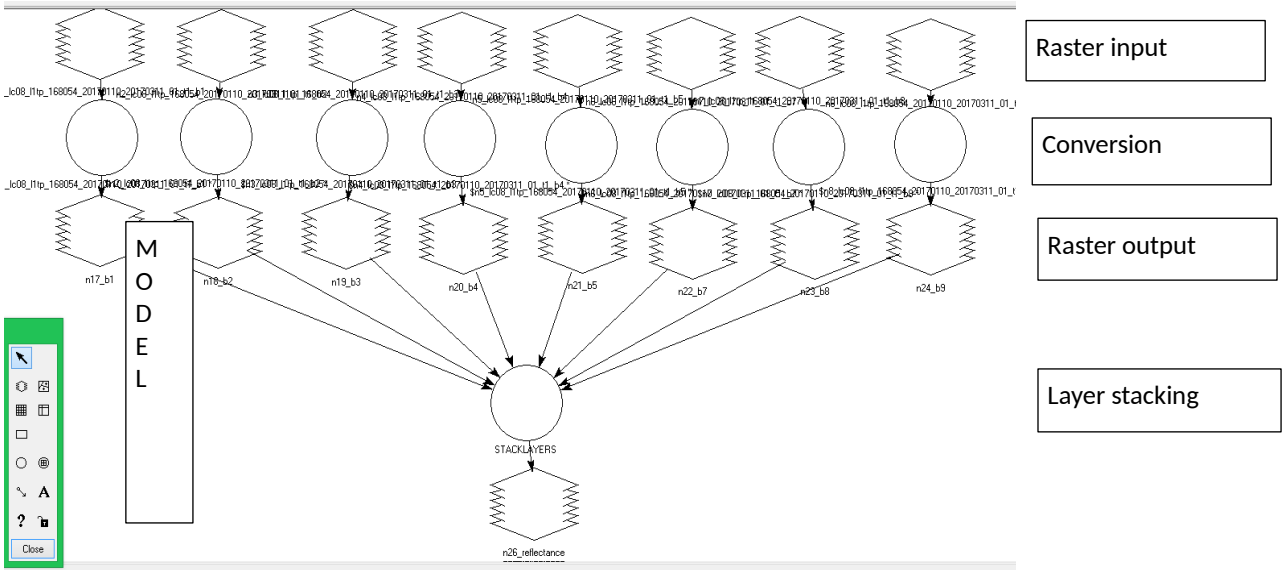

Figure 3:Conversion of Digital Number to reflectance model

Geometric correction:

This step is used to transform remotely sensed image in to a map with scale and projection properties. It is registration of the image to make it usable with other maps or images of the applied referenced system. It can be used in one of the following circumstances:

- ✓ To transform an image to match a map projection
- ✓ To locate points of interest on a map and an image
- ✓ To bring adjacent image into registration
- ✓ To overlay images and maps within GIS

All these satellite images are already geometrically corrected. So, we haven't go further process for geometric correction.

Atmospheric correction

We have applied image processing techniques in order to remove atmospheric effects and to get correct reflectance spectra. This correction is done mainly if there is cloud coverage of the atmosphere during acquisition. All the Landsat images are acquired in winter season. During the data acquisition period the season was atmospherically corrected. So, we haven't done any process for removing the cloud. To improve the semi-automated image preprocessing, we have used the R programming to perform the pre-processing tasks

R code for Radiometric correction and Topographic Correction:

```
##Importing Band
B2_input <- brick ("/Users/Desktop/R_Files/Corrections/B2.TIF")
 ##Digital Number to Radiance
B2_ref <- 0.00002  * B2_input -0.100000
writeRaster(B2_ref,filename = "B2_REF.tif",overwrite=TRUE)
##Topographic Correction
sun_Elevation <- B2_ref/1.09203250728462
plotRGB(sun_Elevation,1,1,1,stretch="hist")
writeRaster(sun_Elevation,filename = "Blue_10.tif",overwrite=TRUE)
##Importing Band
B3_input<-brick "/Users/tagel/Documents/R_Files/Corrections/B3.TIF")
##Digital Number to Radiance
```

Figure 4: Radiometric correction and Topographic Correction

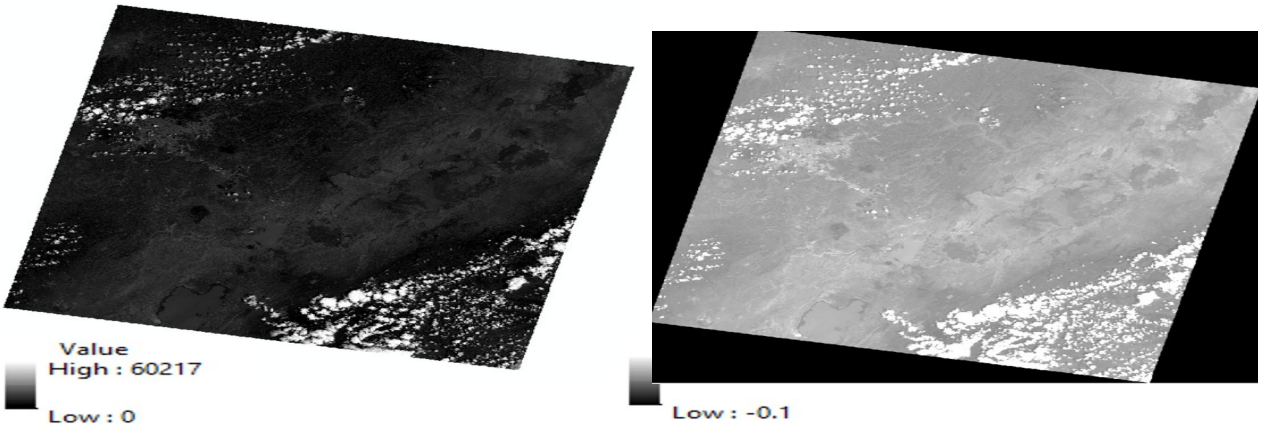

Figure 5:Raw DN Pixels and Corrected Image with Surface Reflectance

After performing the radiometric correction, the dark images are changed to relatively visible. The radiometer correction changes reflectance values but still it not ready to perform the classification task. The next step was to clip the study area using prepared shape-file which shown on Figure 6 below.

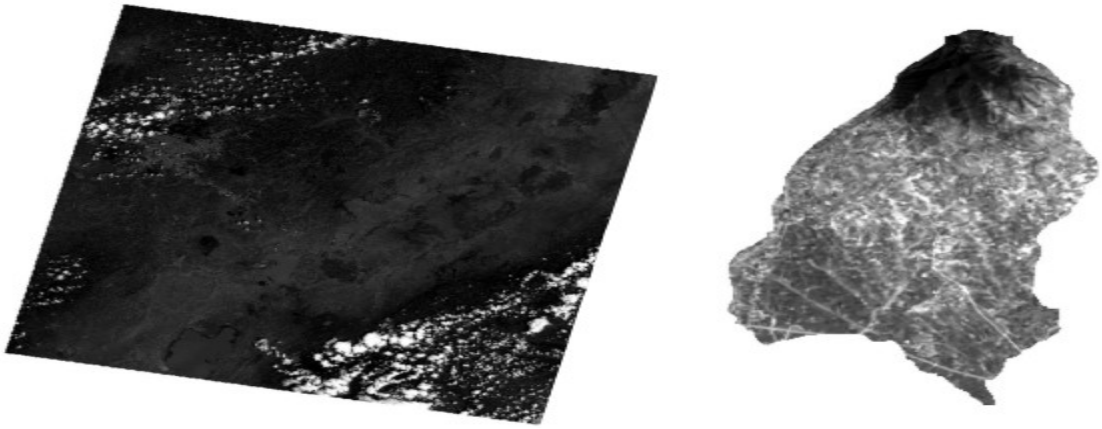

Figure 6:Extracted Study Area from the Tile

From figure 6 above, the clipped image file was a combination of multiple bands. All the spectrum are not utilized to perform land-use land-cover analysis. Therefore, at this stage we have to decide which bands are relevant for our object. Once, we completed band selection, the next step was combining only the selected bands using the image stacking techniques.

The following R code has been used for layer Stacking purpose

```
A=stack(c("Blue_10.tif","Green_10.tif","R_10.tif","NIR_10.tif"))

writeRaster(A, filename="LS_10.tif")
```

ands to obtain the

olor or false color

The next step required a decision to determine the number of bands need for land cover analysis. Based on our objective, we have selected only four spectral bands to extract representative features to characterize objects. So, the final stacked multi-spectral image is the

combination of red, green, blue, and near-infrared bands respectively which is depicted on figure 7 below. For a better visualization purposes, we add different types of false colors.

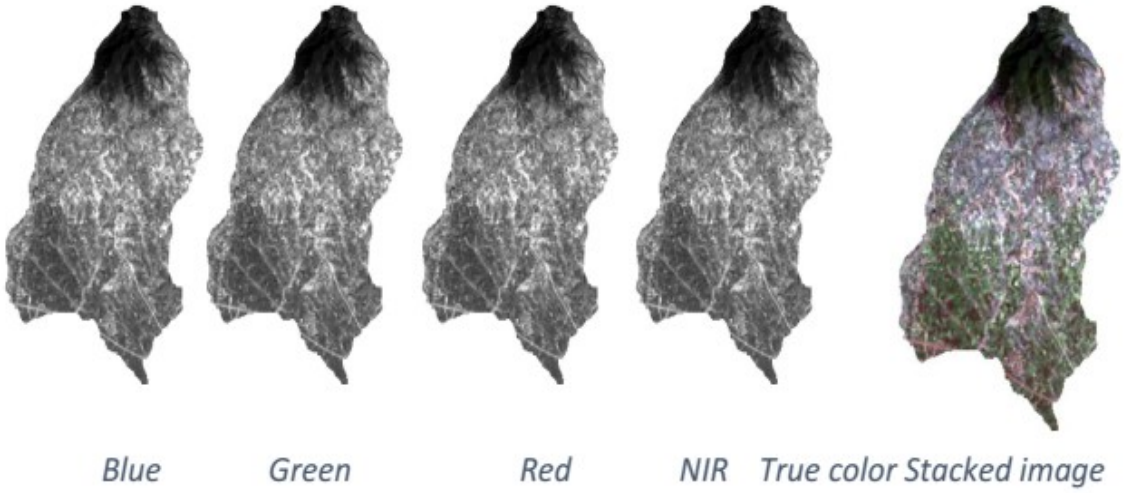

Figure 7:True color stacked image

Finally, the stacked multi-spectral have been used to as input dataset to performance the remaining image processing task. As we can see from figure 7 above, the stacked image represents our study area (Yerer Sillasie).

CHAPTER 4: PROPOSED APPROACH

Proposed system

After the Landsat preprocessing task was completed and once the required staked image was obtained, the next step was to implement the classification using collected image dataset. To implement the proposed image classification models, we have constructed general architecture as shown on figure 8 below.

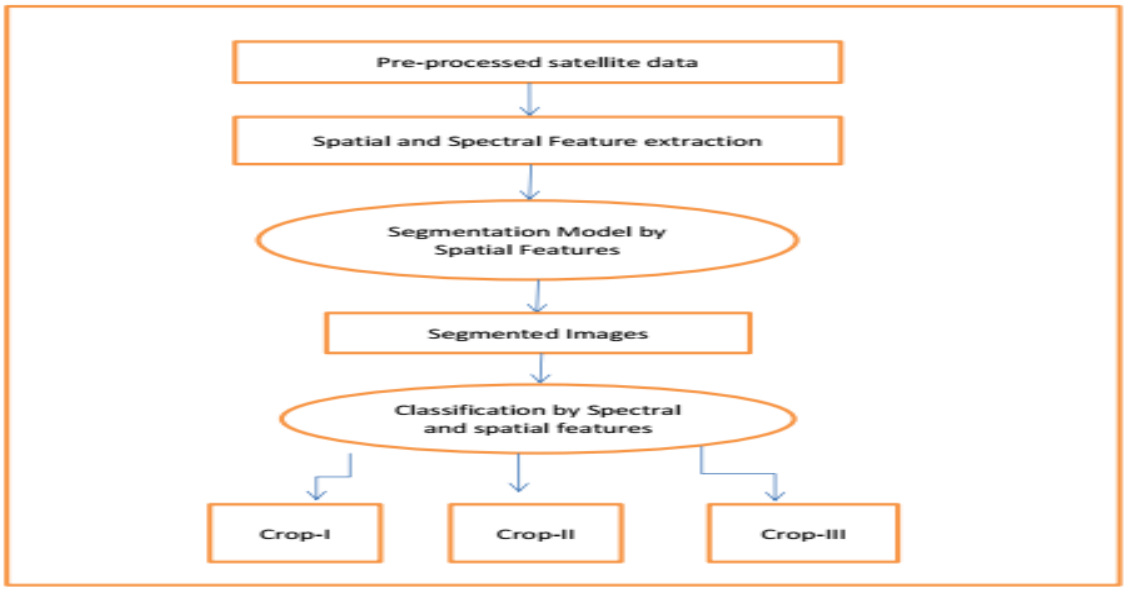

Figure 8: Multispectral Image Processing General Framework

Proposed Algorithm: Steps

1. **Data Collection and Preprocessing Step:** data collection and pr-processing step was a time consuming and laborious task to get the relevant training dataset.
2. **Image** Preprocessing **step:** This step includes band identification and stacking selected bands together. In the case Satellite image data, pre-processing task incorporate various correction activity.
3. **Image Segmentation and classification Step:** The first step was labeling sample training data by identifying the area of interest (AOI). This step includes modeling training for image classification which is used to identify land-cover land use on the study area. We have used sample training dataset to build multiple algorithms to the Landsat image data into the target classes.

4 **Yield estimation Modeling:** Estimate the amount yield in a certain farming field will gives significant insight to make multiple analysis and strategic decision in the domain area. In this research, out main objective was to address the limitation of convention yield estimation process. But, one of the big challenges was the spatial resolution of image data we have used to perform the classification task.

5 **Evaluation of Models:** The performance individual model measured against the ground truth knowledge. We have used confusion matrix to measure the classification accuracy.

Models' Description

To implement the experiment, we have selected four machine learning models. The first two model are QGIS built-in models and the remaining two models selected based on their performance to handle multi-spectral image data. In this subsection, the mathematical equation of each model has been discussed as follows:

Minimum distance classification

The minimum distance classifier is used to classify unknown image data to classes which minimize the distance between the image data and the class in multi-feature space. Minimum distance algorithm uses the mean vectors for each class and calculates the Euclidean distance from each unknown pixel to the mean vector for each class. The pixels are classified to the nearest class. The distance is defined as an index of similarity so that the minimum distance is identical to the maximum similarity. The following distances are often used in this procedure.

1 **Euclidian distance:** Is used in cases where the variances of the population classes are different to each other. The Euclidian distance is theoretically identical to the similarity index.

2 **Normalized Euclidian distance:** The Normalized Euclidian distance is proportional to the similarity index:

$$d_k^2 = (X - \mu_k)^t \sigma_k^{-1} (X - \mu_k)$$

3. Mahalanobis Distance: In cases where there is correlation between the axes in feature space, the Mahalanobis distance with variance-covariance matrix, should be used in performing classification with following equation:

$$d_k^2 = (X - \mu_k)^t \ \Sigma_k^{-1} (X - \mu_k)$$

where **X** : vector of image data (n bands)

X = [x1, x2, ... xn]

μk : mean of the kth class

μk = [m1, m2, ... mn]

The Minimum distance-based classifiers have a tendency to do misclassification if the two classes have very similar values of features.

Maximum likelihood method

The maximum likelihood method (MLE) method used for image classification is based on following likelihood function

Lk = P(k/X) = P(k)*P(X/k) /ΣP(i)*P(X/i)

Where; P(k) is the prior probability of class k, P(X/k) represents the conditional probability to observe **X** from class k, or probability density function. Usually P(k) are assumed to be equal to each other and ΣP(i)*P(**X**/i) is also common to all classes. Therefore, Lk depends on P(**X**/k) or the probability density function. For mathematical reasons, a multivariate normal distribution is applied as the probability density function. In the case of normal distributions, the likelihood can be expressed as follows.

$$L_k(\mathbf{X}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(\mathbf{X}-\boldsymbol{\mu}_k)^t\right]$$

where n: number of bands, **X**: image data of n bands,

Lk(**X**) : likelihood of **X** belonging to class k

μk : mean vector of class k

Σk : **variance-covariance matrix** of class k

$|\Sigma_k|$: determinant of Σ_k

The maximum likelihood method has an advantage from the view point of probability theory, but the following challenges are inherent with this method:

- ✓ Sufficient ground truth data should be sampled to allow estimation of the mean vector and the variance-covariance matrix of population.
- ✓ The inverse matrix of the variance-covariance matrix becomes unstable in the case where there exists very high correlation between two bands or the ground truth data are very homogeneous. In such cases, the number of bands should be reduced by a principal component analysis.
- ✓ When the distribution of the population does not follow the normal distribution, the maximum likelihood method cannot be applied.

Random Forest Algorithm

Random Forest [14] is an ensemble-based machine learning algorithm approach to create a bunch of decision trees with a random subset of the data. First, in which 'n' random trees are created, this forms the random forest. In the second stage, the outcome for the same test feature from all decision trees is combined. Then the final prediction is derived by assessing the results of each decision tree or just by going with a prediction that appears the most times in the decision trees.

Random forest has a tendency to miss-classify and over-fitting. Therefore, the training data should be properly labeled and managing outliers from the dataset. Also, they tend to favor the category in which there are a greater number of levels present in case of categorical classification, therefore the number of levels should be kept even across all the categories.

Let given training set as $\mathbf{X=\{x_1, x_2, \ldots..x_n\}}$ and $\mathbf{Y=\{ y_1, y_2, \ldots..y_n \}}$ be the corresponding response variable, then

A random forest works on following sampling plan:

1 Sample with replacement n training examples from X, Y and call them as Xa Ya

2 Train the tree classifier on the sampled training data Fa

After training several instances of Fa, the final hypothesis can be represented by following formula

$$\mathbf{F=1/A}\sum Fa(Xi)$$

RF [14] is advantageous as its non-parametric nature is suited to remote sensing data and it is simple to train and tune. However, the split rules determined for classification are unknown

and the classification accuracy can be poor when applied to large-scale land cover and land use mapping based on Landsat data.

Support Vector Machines

The final model utilized was Support vector machine algorithm, which is based on the principle of maximum margin and intuitively works on creating linear decision boundaries to **classify** multiple classes [15]. Depending upon the type of data SVM can be used with a linear classification function or highly sophisticated kernel functions can be used in case of nonlinear data [16]. In SVM, a separating hyperplane can be represented by following formula:

$WX+B=0$

Where X is the n dimensional feature [7] vector B is the bias term, W is the weight matrix which has to be learned subject the constraints specified in terms of margin of the separation. It gives two equations for actual classification of the data as follows:

$W1x1+w2x2\ldots\ldots\ldots\ldots\ldots\ldots..wnxn+b>0\ forpositiveY$ and

$$W1x1+W2x2\ldots\ldots\ldots\ldots..WnXn+b<0\ fornegativeY$$

While optimizing for the margin of separation, this on transformed representation takes following form

$d(X')=\sum yi\alpha iXi\ X'+b\ 0/\square$

the above SVM is a linear model which is not as per the requirement of remotely sensed data, because the selected images involve a highly non-linear decision boundary. Therefore, this research has used kernel method with Gaussian kernel. The kernel functions like linear kernel, polynomial or Gaussian kernel can be used to learn a decision boundary in case of non-linear data**:**

Polynomial kernel: $K(Xi,Xj)=(Xi.Xj+1)^h$

Gaussian kernel: $K(X_i,X_j)=e^{-|Xi-Xj|.\vee\ Xi-Xj\ \vee 2\sigma^2}$

Model development

Image segmentation

In this research, the hyperspectral image has been segmented based on the similarity measurement of pixel value to represent the object. Then the objects are characterized in both the raster and vector domains. The objects are classified using both spectral and spatial metrics. Our main objective is to assign all pixels in the image to particular segment.

In this study we have utilized Erdas imagine 2015 and QGID modelling tools to perform the segmentation task. R programming language has been used to write the script to segment and extract similar pixel value in the same segment. R language contain a massive support packages and libraries perform hyperspectral image processing task.

Feature engineering

Feature engineering plays a significant role in hyperspectral image classification. In this research, we have extracted the spectral signature of homogeneous training sample to build the classification model. We have used both semi-automatic and automated featuring extraction techniques using Erdas Imagine and R language respectively.

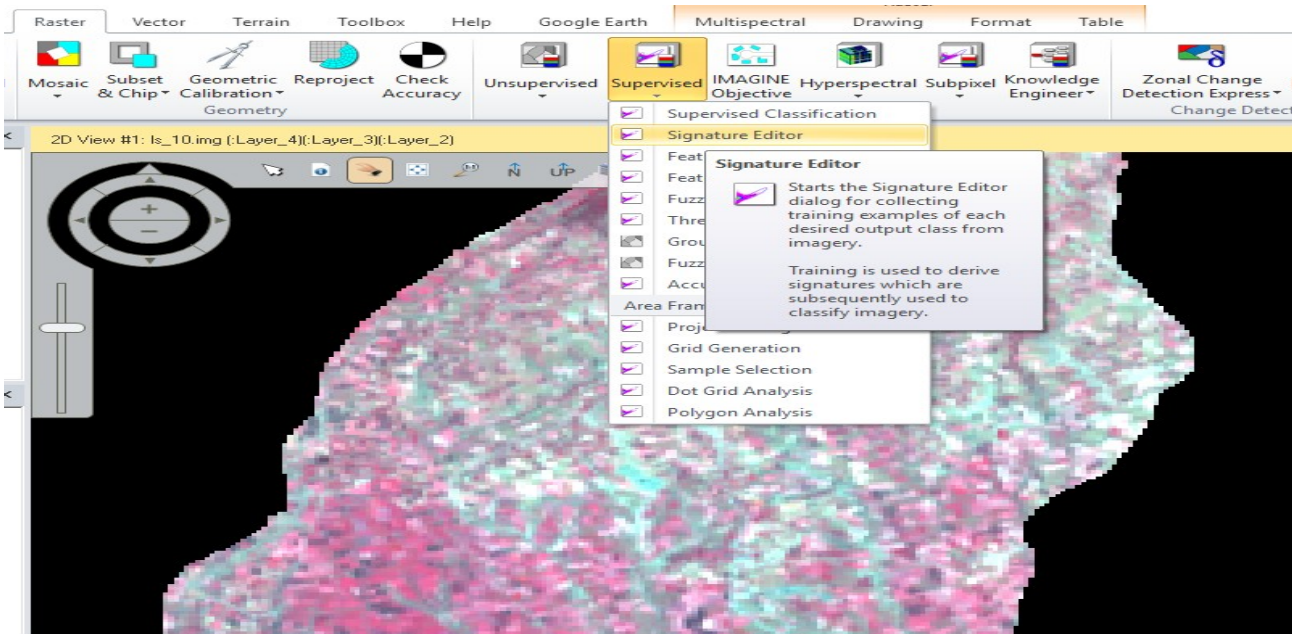

Figure 9:Semi-automated image segmentation process

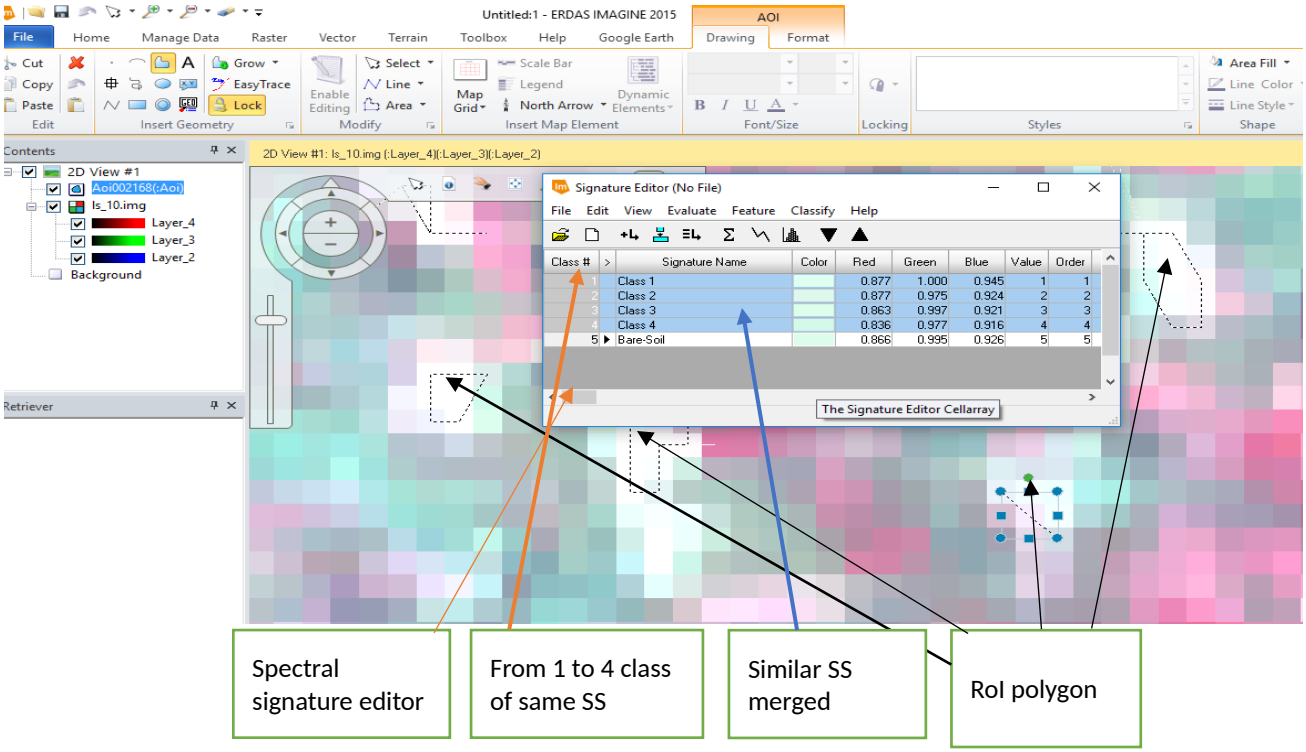

Figure 10:Feature Extraction and Labeling using ROI

The spectral and spatial feature of the hyperspectral [16] image has been carried out as it depicted in the above figure. The spectral signature of each homogeneous training sample has been identified from each region of interest to training the classifier (ROI). Determining the ROI requires the support of domain expert from the study area. We have used the spectral signature of each ROI to classify the imagery and all homogeneous ROI can be merged into single class have the similar spectral pixel value. On the other hands, training sites uses geometric tools to draw the polygon to extract the spatial features (geometric coordination, pixel size of the polygon, density, grid box etc.) of the image data. A Supervised classification approach is appropriate when you want to identify relatively few classes, you have to select training sites that can be verified with ground truth data.

Extracting training pixels values

Extract the pixel values in the training areas for every band in the Landsat image and store them in a data frame (called here dfss) along with the corresponding land cover class id**:**

```
dfss = data.frame(matrix(vector(), nrow = 0, ncol = length(names(img)) + 1))
for (i in 1:length(unique(trainData[[responseCol]]))){
  category <- unique(trainData[[responseCol]])[i]
```

```
  categorymap <- trainData[trainData[[responseCol]] == category,]
  dataSet <- extract(image, categorymap)
  dataSet <- lapply(dataSet, function(x){cbind(x, class = as.numeric(rep(category,
 nrow(x))))})
  df <- do.call("rbind", dataSet)
  dfss <- rbind(dfss, df)
}
```

Classification

we trained our machine learning models using the extracted features to classify the entire Landsat image data. The target classes comprise different types of land-covers. In this study, We have used about 30 meters spatial resolution satellite image data. Due to the image resolution constraint, it difficult to conduct classification at crop level. In the current paper, we computed the land use land cover analysis. The following table describes the target classes of each individual land usage to implement supervised classification task.

Table 3:Land cover Land use and its description

Classes	Description
Built up	Includes all residential, commercial, and industrial development.
Water body	It's any significant accumulation of water.it refers to oceans, rivers, streams, canals, seas and lakes
Bare land	Bare land or most often representative of bare earth or soil. Land areas of exposed soil and barren area influenced by human Agricultural land and grass land
Dense vegetation	An area dominated by plantation
Sparse vegetation	It's a plant community characterized by vegetation dominated by shrubs.

Once the target classes have been determined, then we utilized different false colors to represent each target classes.

Setup experiment

For the purpose of classification tasks, we have used three different satellite images collected during the period of 2018-10-12, 2018-11-12 and 2018-12-12 from the same area. During data pre-processing, all the required correction has been made to obtain the final raster image data. Then geometric, atmospheric and radiometric correction has done. In addition, the relevant bands has been selected to characterize objects for the purpose of land cover analysis. The following figure describes the detail description of our datasets used to training the models.

```
class      : RasterBrick
dimensions : 327, 200, 65400, 4  (nrow, ncol, ncell, nlayers)
resolution : 30, 30  (x, y)
extent     : 492555, 498555, 972435, 982245  (xmin, xmax, ymin, ymax)
crs        : +proj=utm +zone=37 +ellps=WGS84 +towgs84=0,0,0,-0,-0,-0,0
+units=m +no_defs
source     : C:/Users/Tagel/Desktop/R_FILES/SVM_10/ls_10.img
names      :   ls_10.1,   ls_10.2,   ls_10.3,   ls_10.4
min values :         0,         0,         0,         0
max values : 0.1808769, 0.2040188, 0.2481846, 0.5377183
```

The description gives us detail information about the training datasets. Landsat image provide contains rich information about the spatial and spectral feature about each object such as the class type, dimensions of image data, image resolution, numbers of layers, and min or max values of the training data.

The stacked input image data was the combination of four bands namely (green, blue, red and near infrared red) bands. Each training sample contain a pixel value of the above four bands are used represent the respective classes. Some selected band values and their respective target classes have been shown on Table 5 below.

Table 4: Sample training Dataset Extract from each channel

SN	B1	B2	B3	B4	class
1	0.07458609	0.06879496	0.04646420	0.2344842	1
2	0.07456354	0.06715000	0.04592339	0.2124689	1
50	0.07859705	0.07893506	0.05721270	0.2861987	2
51	0.07888999	0.07976880	0.05791124	0.2933869	2
52	0.07839426	0.07600570	0.05444108	0.2989978	2
71	0.10300090	0.12341630	0.11104538	0.4008495	3
72	0.10802589	0.13409720	0.12508379	0.4185608	3

```
 73  0.11616050 0.14164594 0.14347117 0.3987313    3
107 0.12929755 0.13896446 0.16230923 0.2793936    4
108 0.13299307 0.14948763 0.17907420 0.3312433    4
109 0.12422751 0.13727444 0.15449007 0.3581259    4
```

After performing all the required raster image data description and characterization of each band's values (pixel value). The next step was to training our model using the training dataset, in this regard about 147 training datasets have been used. Each training sample is the combination of four bands namely (B1, B2, B3 and B4) to represent image specific vector value for the purposes of this experiment, MLE, MinDis, RF, and SVM model has been used to classify in the image data into the respective categories.

Training the proposed models

Landsat image were imported into R as a RasterBrick object using the brick function from the 'raster' package. Also replace the original band names (e.g., 'LS_10_1') with shorter ones ('B1' to 'B4').

```
img <- brick("ls_10.img")
names(img)<- c(paste0("B",1:3, coll="),"B4")
```

RGB visualization of the Landsat image in R using the plotRGB command, for example, a false color composite RGB 4:3:2 (Near infrared - Red- Green). Using the expression img * (img >= 0) to convert the negative values to zero:

```
plotRGB(img * (img >= 0), r = 4, g = 5, b = 3, scale = 10000)
```

set of training areas in a polygon shapefile ('tls_10_TD.shp') which stores the id for each land cover type in a column in the attribute table called 'class' as shown below: the shapefile function from the 'raster' package to import this file into R as an object of class SpatialPolygonsDataFrame and let's create a variable to store the name of the 'class' column:

```
trainData<-shapefile("ls_10_TD.shp")
responseCol <- "Class"
```

The data frame resulting from working with my data has about 8 K rows. It is necessary to work with a smaller dataset as it may take a long time to train and fit in case of Random-Forests model.

```
nsamples <- 147
sdfss <- subset(dfss[sample(1:nrow(dfss), nsamples), ])
```

Once training sample has been randomly selected, the next was fitting the model using training datasets. We have used ‘RF’ and “SVM” models which stands for the random forest algorithm and support vector machine respectively.

To resolve the computational complexity, we have used the clustering function in r which supports multi-core computing to speed up computation time. We just need to add one line for creating a cluster object and another one for deleting it after the operation is finished. The following scrip how to include clustering method during implementing model prediction.

```
beginCluster()
preds_rf <- clusterR(landSat-img, raster::predict, args = list(model = modFit_rf))
endCluster()
```

CHAPTER 6: RESULT AND DISCUSSION

On chapter five, under experiment setting section, we have discussed about data description, training sample size, and model fitting and predicting the target class. In this chapter, the experiment output from each model has been discussed as follows. To classify the Landsat image data, first have employed RF and SVM classifiers. The following parameters have been used to training the models.

Summary of RF and SVM classification

1. 147 samples are randomly selected to train the classifier
2. 3 predictor
3. 4 classes : '1', '2', '3', '4' were determined for spectral classification
4. Resampling: Bootstrapped (25 reps)
5. Summary of sample sizes: 147, 147, 147, 147, 147, 147 ...
6. Resampling results across tuning parameters

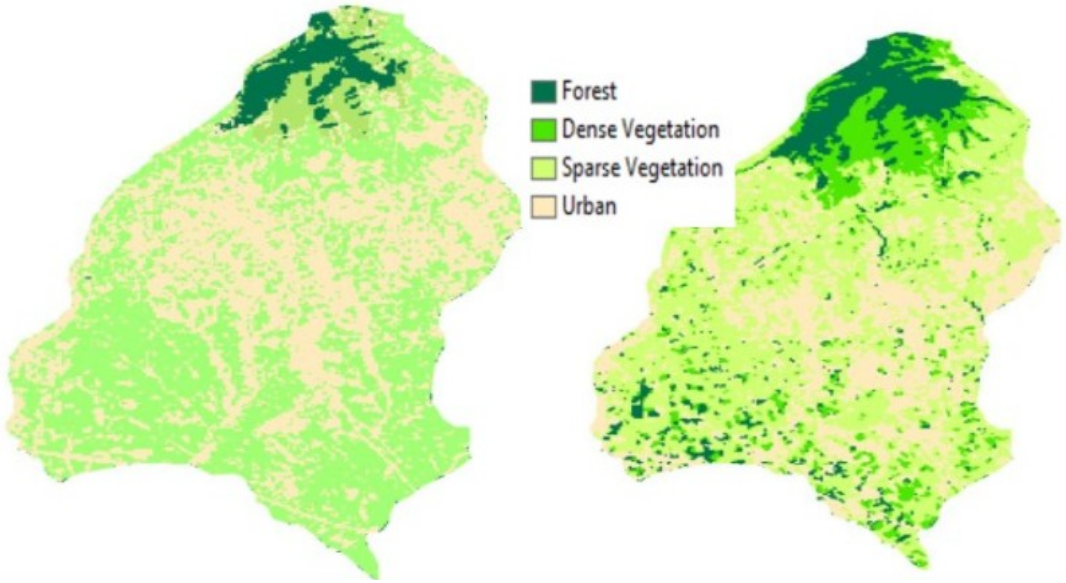

Figure 11:Experiment Result of RF and SVM on 2018-10-12 Image data

Figure 12 above shows the experiment output of both RF and SVM models given parameters. We have used the first image (October 2018-10-12) as input to classify into the respective target classes. During the experimentation, our objective was to visualize the patterns of land-cover between October to December. In these three the pattern of land cover very dynamic the main reason, in this period their massive activities of crop maturity and harvesting in majority of the farming area. The other important issues were, Satellite image data was multi-dimension and complex to handle by classical models. But the experiment result showed that,

the proposed algorithms are competitive enough to handle the multi-spectral image data with best classification accuracy. This is due to the small size training sample we employed to build our model. The experiment out on the training data has been summarized as follows:

Table 5: Summarize of model's accuracy on the training datasets

SN	Model name	S. size	Predictor	Resampling T.	N. class	Optimal value	Acc
1	RF	103	3	Bootstrapped	4	Mtr=2	98.5
3	SVM	103	3	>>	4	C=1, sigm=1.810	97.72%

One of the possible solutions is increasing the sizes of both the training and testing data set. Another challenge is, the sensitive of base classifiers, where small variances significantly affected their prediction performance. We have trained our models for several times until we get the optimal predictor value which affect the performance our model. After a number of trial and errors we set predictor value to be three, where we got optimal classification performance. On the other hand, the final value used for the model was mtry equals to two. Similar procedure has been applied for each based classifier.

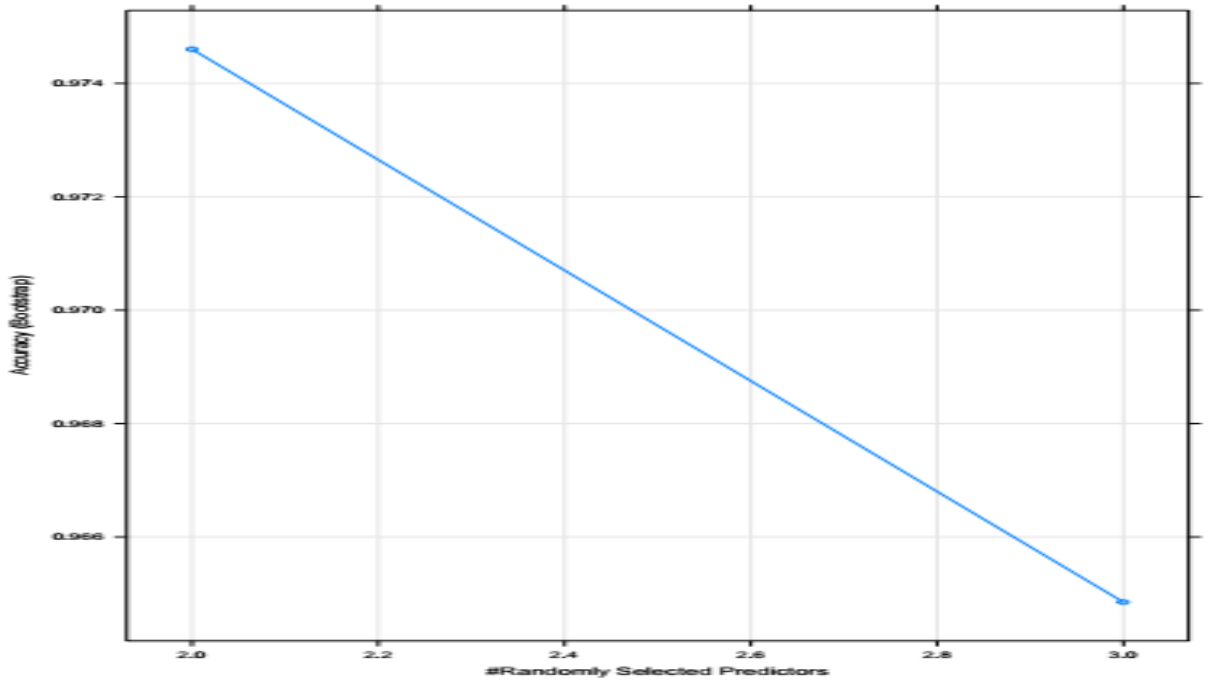

Figure 12:Correlation between classifier's performance and predictor parameters

The experiment output showed that the selected models were promising in classifying multi-spectral image data into the respective categories on training datasets. Then made model performance assessment using test data. Table 6 below summarize the prediction performance of each model using testing datasets.

Table 6: Summary of model's prediction accuracy

Models	Prediction accuracy using testing data	Remark
RF	98.5% and 100%	Model sensitivity
SVM	97.72% and 95.45%	

From Table 6 above, we can see that the performance of each base classifiers are competitive enough to predict on testing datasets. From this result, we can see that RF algorithm is capable enough to predict the data set with optimal classification performance than SVM. One of the big challenges were, classical models are very sensitive. Small change on training sample would bring a significant different on the classification performance of the models. In the case of SVM model, we need to address the non-linearity issue by finding the best fitting kernel function [16][12]. The idea behind generating non-linear decision boundaries is that we need to do some nonlinear transformations on the features Xi, which transforms them into a higher dimensional space. In our case, the non-linear decision boundary and the values of the tuning parameters were c = 1, r = 1.8 and a number of support vectors. Similar model tuning is required for Random Forest classifier to handle the bias-variance trade-off.

Experiment Result using Modelling Tools

In this study, we have made other experiments using geo-spatial modeling tools such as QGIS and Erdas Imagine to classify multi-spectral image data. The modelling tools have a built-in classification algorithm such as Maximum likelihood Algorithm. Feature engineering task have been done using Semi-automatic image segmentation techniques to discriminate one pixel value from its neighboring pixel. The expert’s domain knowledge and image quality and resolution have a significant impact on the classification performance of built-in models. One of the challenges were training samples are manually labeled by creating region of interest using

different polygons. In addition, mixed pixel values degraded the classification performance of each model. Figure 13 below shows the experiment output of Maximum likelihood model.

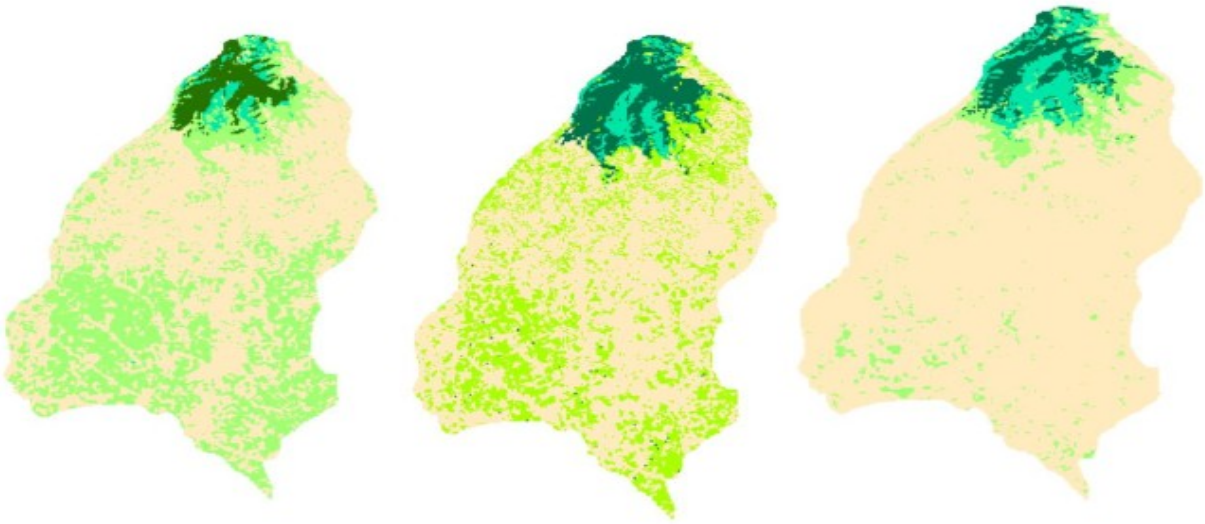

Figure 13:Experiment output of three image data

From the experiment result on Figure 13, we can visualize the patterns of land-cover in period of October to December. Majority of the land in the study was covered by crop and crop harvesting started late November. From the experiment out result most of the farmer completed harvesting the end of December that is why the land cover become bare land. On Table 7 below we have summarized the patterns of land cover from September to December 2018 based on the experiment output shown in figure 12.

Table 7:Land use land cover pattern of the study area

Class ID	Target class	Area (Km^2)		
		2018-10-12	2018-11-12	2018-12-12
1	Forest	2.21	1.4598	1.4355
2	Dense Vegetation	1.5543	1.0476	0.6903
3	Sparse Vegetation	16.3143	12.2598	2.8962
4	Urban	**38.7905**	**44.4501**	**52.9128**

From the table 7 above, the selected study area has a visible pattern of land cover in the given time period. At the early October majority of the land cover by dense and sparse

vegetation and forest. The vegetation cover dynamically decreases by the following month to show that it is harvesting period for many farmers in the study area. From the experiment result, majority of land cover become bare land in the third month, during this period many harvesting processing completed by many farmers.

Classification Accuracy using confusion matrix

The performance of MLE algorithm has been evaluated using confusion matrix. Performance assessment are done on the basis of how many of similar pixel values are classified in the target and how many of the pixel values are wrongly classified into other classes. Table 8 below summarizes the classification performance of MLE model using confusion matrix.

Table 8:Summarizes of confusion matrix

Data	Sparse Veg	Dense Veg	Forest/Dense Veg	Urban
Sparse Veg	48	0	17	0
Dense Veg	0	25	0	0
Forest/Dense Veg	0	0	163	0
Urban	0	0	0	98
Column Total	48	25	180	98

Data	Forest / De	Sparse Veg	Dense Vege	Urban
Forest/ De	530	0	0	0
Sparse Veg	0	38	0	2
Dense Vege	0	0	10	0
Urban	9	2	0	44
Column Total	539	41	10	46

Data	Forest /Den	Dense Vege	Sparse Veg	Urban
Forest/Den	534	0	0	0
Dense Vege	4	39	0	0
Sparse Veg	0	0	25	0
Urban	0	0	0	53
Column Total	538	39	25	53

Accordingly, the classification accuracy of MLE for each multi-spectral image datasets were 95.15 %, 96.8% and 99% respectively. Similarly, the experiment result shows that the MLE algorithm is potential capable to classify the image dataset. But manually labeling training sample was prone to error. Data quality and spatial resolution of multi-spectral image data was another challenge in the domain area. To handle the limitations; first employing automatic feature extraction techniques would improve the classification accuracy. Second, acquisition of higher spatial resolution satellite image data would enhance the classification accuracy of the models. Finally, we have made the validation of proposed model using overlapping techniques of training shape files on the google earth image with the respective period. The spatial correlation was examined visually. The following figure has been used for the purposes of

validating the proposed model to classify land cover of study area using image data from the google earth.

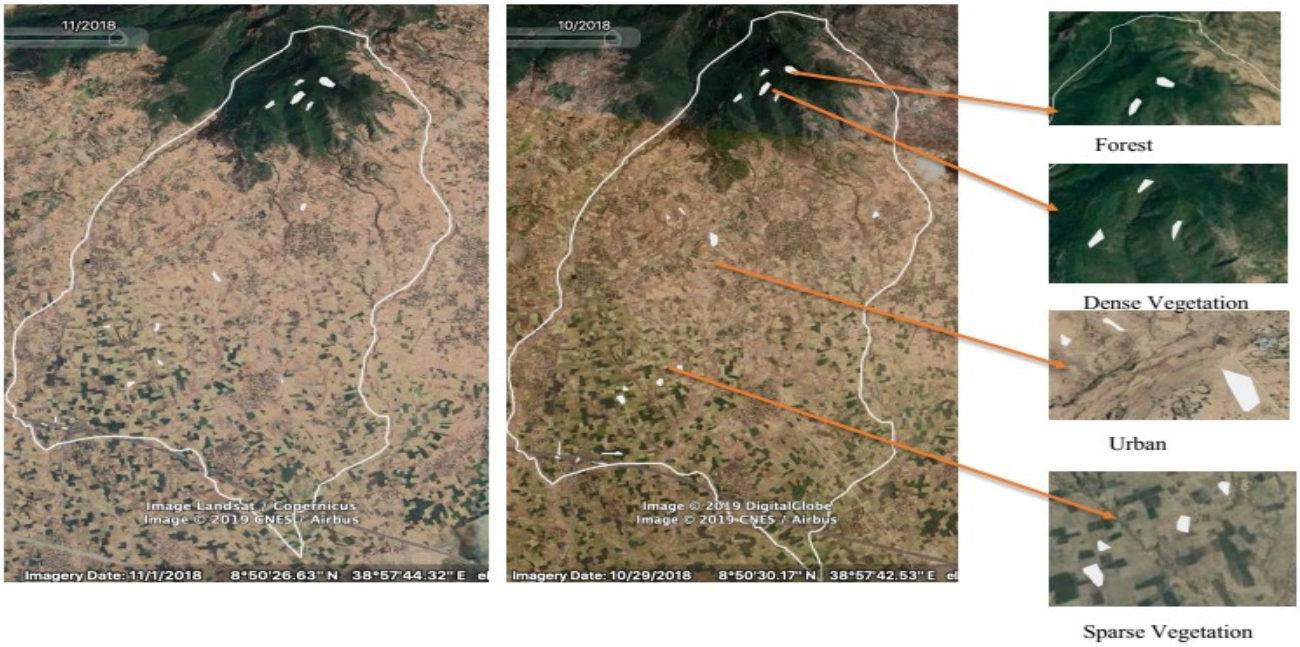

Figure 14:Google Earth image data for validation purpose

YIELD ESTIMATION MODEL

The second part of this study focuses on crop yield estimation [4] using multi-spectral image data. In case of Ethiopia, yield estimation was done by survey farming field and collecting sampling crop to compute yield for the total area. This convention method was prone error due to multiple factors such as; sample representatives' problems, there is no uniformity of crop production in different ecological zone, and difficulty of generalization. In this research study we have considered different indexing measurement to analyze the crop health status. The analogy was if the crop is health, then the yield estimated to be higher. Vegetation Indices (Vis) [10] obtained from remote sensing-based canopies are quite simple and effective algorithms for quantitative and qualitative evaluations of vegetation cover, vigor, and growth dynamics, among other applications. In this section we have discussed different indexing measurement used to estimate crop yield.

NDVI

NDVI is calculated from measurements of percent reflectance (p) in the red and near-infrared (NIR) region of the electromagnetic spectrum. Percent reflectance is the between up-

welling (from the canopy) to down-welling (from the sky) radiation. Requires a measurement of both. Red bands is related to chlorophyll content (high absorption), NIR band is related to leaf cell structure (high scattering).

NDVI = PNIR – Pred / PNIR + Pred, where NDVI values range -1 to 1. There are a number of indexing measurement used to conduct analysis on crop diseases. In this sub-section, we discuss some of the main indexing measurements:

Lear Area Index

NDVI is frequently used to estimate LAI in time (e.g. deciduous canopy) and across space. NDVI 'saturates' when LAI exceeds ~ 3-4 m2. Crop LAI map derived from NDVI image reveals spatial heterogeneity in LAI within and among management units.

Light Interception

Know fraction interception of light (fPAR) is needed to estimate how much light is being captured by a plant canopy (structural photosynthetic capacity). Light capture can be used to predict canopy productivity. A majority of light capture and thus, photosynthesis occurs within the topmost layers of most plant canopies. NDVI estimate light interception is not as prone to saturation compare to LAI estimation.

Phenology

Tracking plant growth and seasonal canopy development (phenology) in annual or deciduous canopies is a common way to use NDVI. Curves tilted to annual or inter-annual NDVI time series can be used to estimate the timing of phonological events such as bud burst, onset of seasonal carbon uptake, and Senescence. Accurate phenology data are a critical component of ecosystem models.

Canopy productivity

In ecosystems with strong seasonality in LAI, NDVI time series can be useful in estimating the length of the carbon uptake period during the growing season. For example, annual grass lands typically have short but pronounced growing season where seasonal productivity is strongly dependent on LAI. The color, texture, size, tone and other spatial features were used to discriminate the different between the types of crops in the selected study area.

The visual interpretation in case of conventional method was subjective and biased to get the prices estimation of crop yield in a certain farming field. In addition, field survey method was subject to many barriers such as time constraining, representative issues, dependency on expert

judgment, data process and interpretation challenges and fabrication of false report are the main pitfall in the domain area.

In this research study, we proposed automated yield estimation approach to mitigate the limitation of field survey method and dependency of yield estimation on indexing measurements. Figure 14 below shows the proposed architecture for yield estimation system.

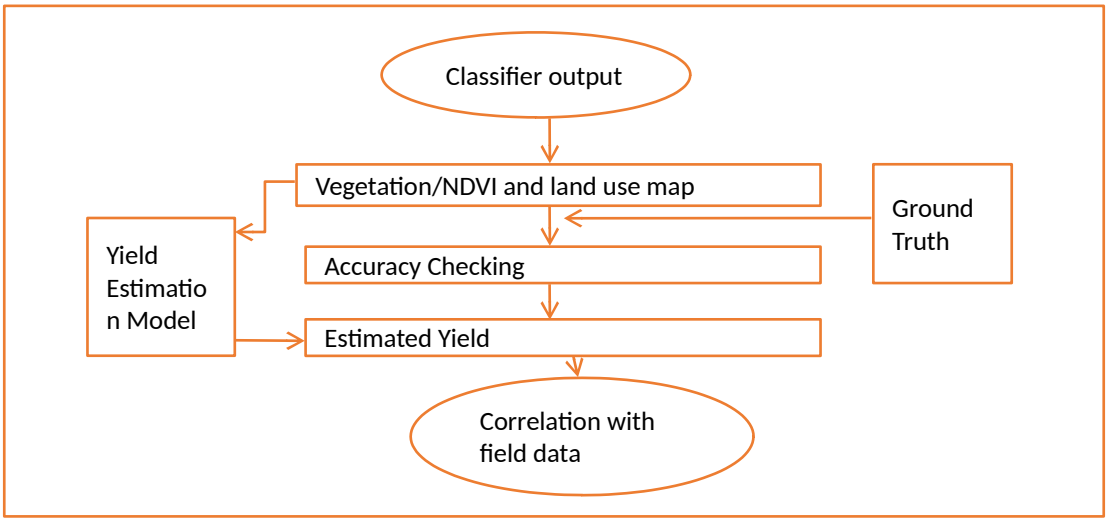

Figure 15: General architecture for yield estimation

Yield of a crop is directly proportional to the vegetation index of the crop under study. The vegetation index is computed based on the reflectance of electromagnetic waves of the crop. The good reflectance crop leaf indicates the healthy status of the crop and a means to estimate productive of the crop on a certain specific area. In this research, we have used Normalized difference vegetation index also called NDVI values to calculate the yield. The Normalized Difference Vegetation Index (NDVI) is a measure of the difference in reflectance between these Wavelength ranges. NDVI takes values between -1 and 1.

NDVI = PNIR – Pred / PNIR + Pred, where NDVI values range -1 to 1. In the case of yield estimation, ndvi spectral reflectance values were used to compute crop health. Healthy crops have higher ndvi value and crops affected by different diseases have different lesser ndvi values. Compare to the conventional field survey techniques, the ndvi measurement would give better insight with rich information about the crop status. But, to provide a quantitative measurement we have explore further in the domain area to provide a accurate yield estimation. One of the

challenges in the conventional method was the process of data collection and analysis took long time and subjective judgment.

Computing the yield

In this study, we have planned to design automated yield estimation method to handle the limitation of classical approaches. But we face a challenge due the spatial resolution of Landsat data utilized for the classification purpose. As we mention in data pre-processing section, our satellite image has 30 meters spatial resolution which is not applicable to compute yield estimation. To resolve the challenge and show the demonstration, we have utilized freely available high resolution image data. The input images are imported into Ecognition developer editors. Once the image has been imported the next step was identify sample area to count the number of crops heads per meter square.

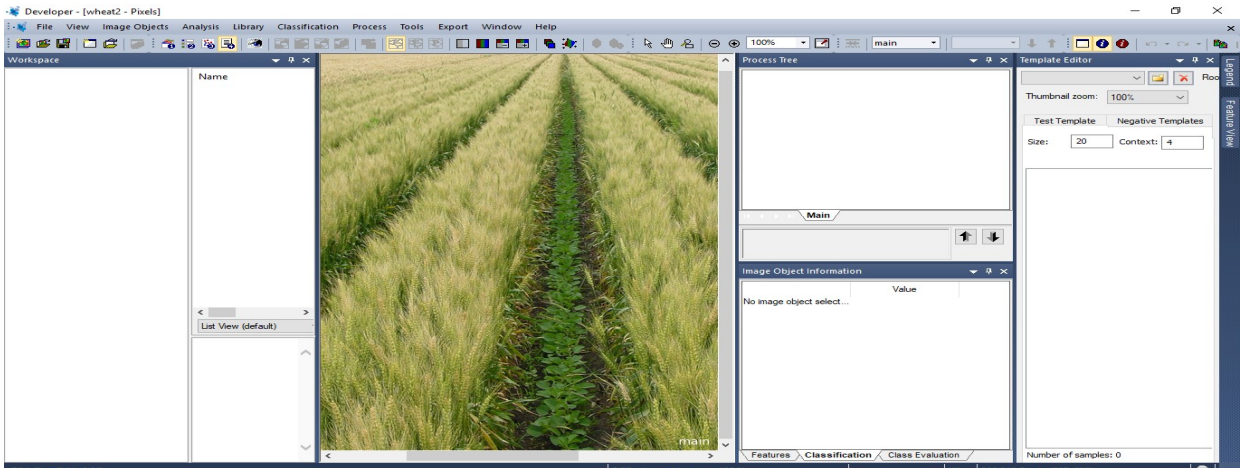

Figure 16:Input image data to compute the yield

The size of crops head was the main limitation in computing yield estimation on a certain specific farming field. After the all the required configuration has been performed, the next task was to select samples of wheat heads/pods to training the classifier based on pixel value similarity. The following figure shows the process of sample selection from the given image data.

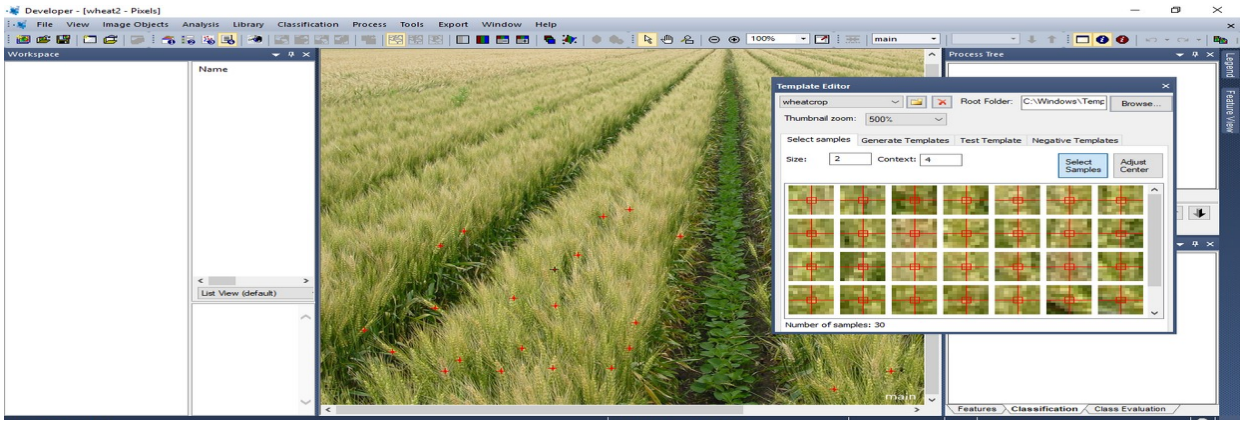

Figure 17:Sample crop head selection steps

After the sample selection procedure has been properly completed, the next step was to perform the classification on the entire image dataset. In this step some critical decisions are required such as determining the numbers of sample size, re-sampling if any, determine the threshold of pixels center and adjusting the threshold and selecting layer group to obtain a good visibility. In this research, we selection a threshold of 0.5 and layer group 1 respectively. The following figure shows the classification process based on the training sample.

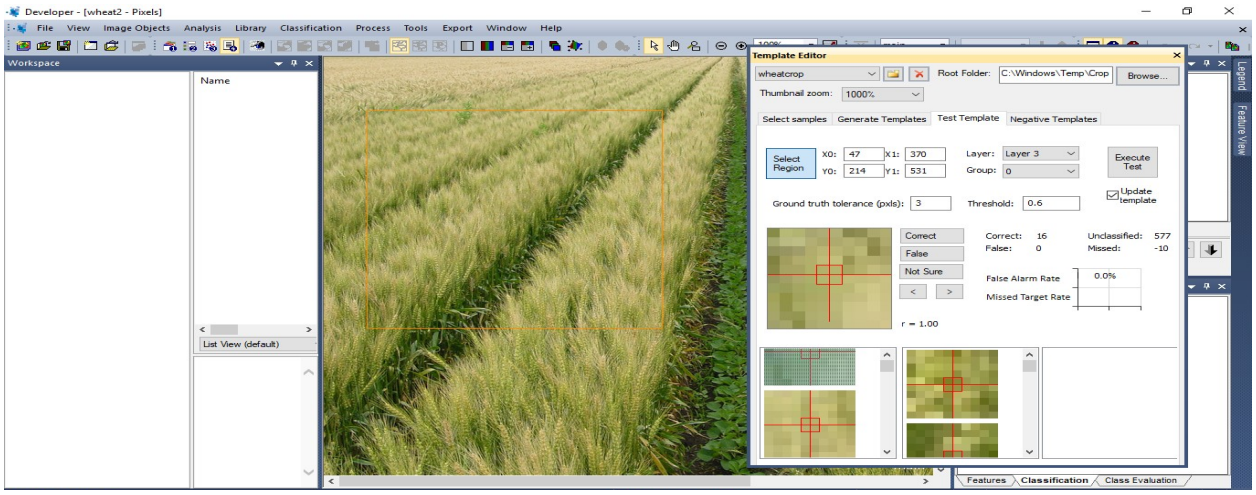

Figure 18: Wheat Crop head classification process

The next subsection is to create template matching of a sampled crop head to evaluate with the ground truth. The following figure display the procedure of template creation process.

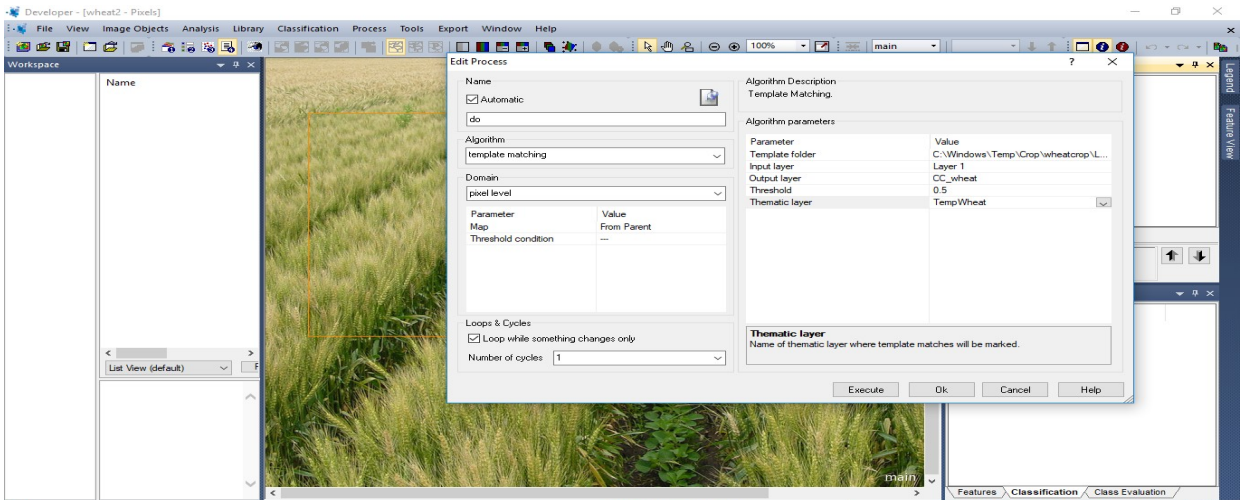

Figure 19: Creating template matching for crop head

Finally, we have imported the classified image to Arch Map10.4 to count the numbers of wheat heads. In addition, we have made frequent discussion with domain expert to get the ground truth information with predicted result. One of the big challenges was the size of crop/wheat head or pods is too small to visualize on the arch map. This makes visualization of numbers of wheat head or pods per/square very difficult from the image data. but the method was efficient enough for other plant such as crops and fruits to calculate the number of yields in a certain specific farming area. This method significantly supports the domain expert to process large size data in short period of time with accurate outputs.

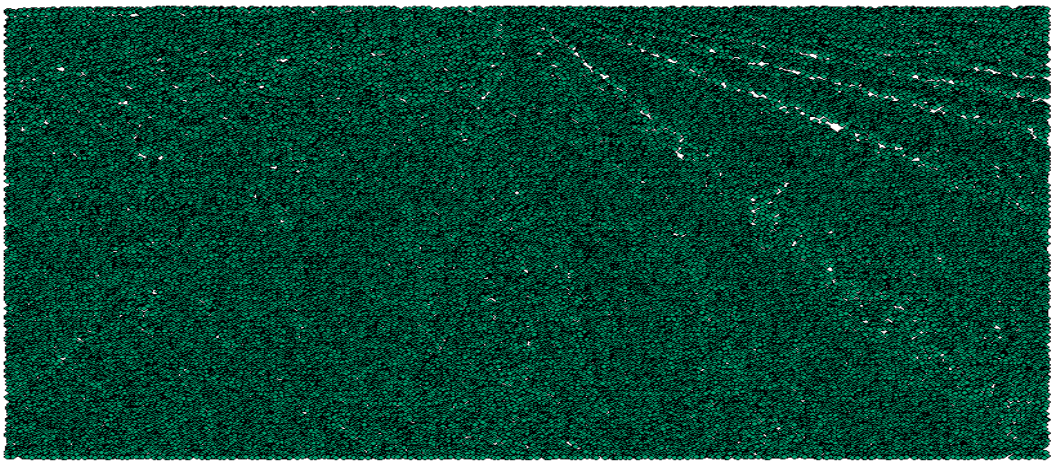

Figure 20: Template matching figure on Arch map10.4

YIELD ESTIMATION PROCEDURE

In this research, the final was estimating crops (wheat) yield using multi-spectral image data to improve prediction accuracy. Once we the get the numbers of pods per square meter then the remaining would be very easy. Accuracy of yield estimates depends upon an adequate number of counts being taken so as to get a representative average of the paddock. The yield estimate determined will only be a guide and assumptions made from the estimates contain a degree of uncertainty. This type of yield estimation is one of the easiest and quickest to complete and should be able to be used in a number of situations on a grain growing property. We have utilized already the defined parameters to calculate the yield estimation and Table 9 below summarized the process.

Table 9:Methodology for estimating wheat yield

Parameter	Lebel	Crop Types	Yield Estimation Procedure
Number of heads/pods per square meter	A	Wheat	250
Average number of grains per head/pod	B		58
Number of grains per square metre = AxB	C		= 250 x 58 = 14500
Yield per square meter = C/100 x 3.4gms	D		= 14500/100 x 3.4 = 493.00gms
Yield in t/ha = D/100			= 493/100 = 4.93t/ha (44.72 quintal)

From the above table, the yield estimation result shows that nearly 44.72 quintal per hectare have been obtained. These results are affected many factors such as soil property, types of seeds, water content, fertilizer inputs, weather conditions and many others. In this research our main focus was to analyze the ndvi values and resolution of image data to process the yield per square meter. Designing automated yield estimation system would improve the process of market analysis, strategic planning for harvest, drafting strategy for import, and demand and supply analysis purposes.

CHAPTER 7: CONCLUSION AND RECOMMENDATION

Conclusion

Currently, hyperspectral image classification has been playing significant role in the domain of digital agriculture. Landsat image data has been used to identify land use land cover (LULC) is one of the hot research areas that needs further exploration in the application of advanced machine learning methods. Landsat image data provides rich information to characterize object. One of the application areas was improving precision agriculture using hyper-spectral or multi-spectral image data.

The main challenges in Ethiopia agricultural sector were the methods of data processing approaches to feed the decision makers. Many of the stakeholder's process data manually which high prone to many errors, amount time it takes, getting representative sample collected using a survey approach, domain expert data interpretation skill, absence of technology in the sector, false report by data collectors, and other factors that affects the sector. Due to these reasons, it is very difficult of analyzed the land cover land use in certain area and specifically it is hard job for farmer and stakeholder to estimate the yield. Accurately estimated yield information on specific farm land has multi-dimensional significant impact in the domain area, decision maker, market analyst, strategic planner and so on.

On the other hands, for the last few decades, machine learning dominating a number of different research and application domain areas. Applying machine learning method to improve precision farming using hyperspectral image dataset is the hot research area. In the case of Ethiopia, few research has been reported in the area of hyperspectral image processing on some crop such as coffee. Alleligni from Bahir Dar University conducted research on crop disease identification using ANN algorithm.

In case of Ethiopia agriculture is the bedrock to maintain sustainable economic development. It plays a key part in long term economic growth and transformation. Due the rapid global population growth, it would be very difficult for many developing countries to address the food security issues. One of the possible solutions would be to automate the process of crop management and adoption of innovative technology in the agriculture sector. In this research, we have an attempt design machine learning model for the purposes of land-use land-cover analysis and yield estimation using multi-spectral image data.

An experimental research methodology has been utilized to perform the classification task. Landsat image data has been collected from Yarer Selassie which is located in Bishoftu

Oromia region of Ethiopia part of the east Shewa zone in the great rift valley at 38°57’40.863"E and 8°50'55.016"N. Yerer Selassie bordered on the southern part by Dugda Bora, on the west side by the West Shewa Zone, on the northwest side by Akaki, on the northeast part by Gimbichu, and on the east by Lome.

During raster image pre-processing, all the required image correction and feature extraction task has been made to get the final stacked image to conduct the experiment. Once the stacked image was obtained, the relevant pixel value has been extracted from the raster image data. Landsat image data processing was a challenge task for many researchers due to its higher dimensionality, Hughes phenomenon due to unbalanced training samples, poor spatial and spectral resolution of image data, larger size of spectral features, and presence of mixed pixels. To training the model, maximum likely hood (MLE), random forest (RF), support vector machine (SVM) algorithm has been utilized to classify the LandSat image data into the target classes. MLE model was built-in algorithm and integrated within QGIS, ArcGIS geo-spatial modeling tools. In the case of modeling tools feature are labeled manually by the domain expert. Whereas, in the case other models (RF and SVM) the processing of feature extraction and training sample selected have been done using R programming scripts.

From the experiment results, we have seen that the method is capable model to handle complex hyperspectral image data. Properly feature extraction and labeling representative training dataset would improve the classification performance of each model. In addition, we have handled the challenges of model over-fitting.

In this research, the main purpose was to handle the limitations of conventional yield estimation approach followed by CSA and the agricultural sector. Ecognition modelling tool and R programming has been used to perform the yield estimation task. Yield estimation using remotely sensed image data significantly address the bottlenecks mentioned in the statement of the problem in the domain area. To implement yield estimation using satellite data, it demand the acquisition of high resolution image data. Another challenge faced were the problem of counting the number head per square meter due its size. For small head sized crops it is difficult to count the numbers of head/pod per square. But it works efficiently for medium and land large sized crops and fruits. Generally, this researcher study would have the following contribution for the end user and scientific community.

1. We have implemented a machine learning approaches to handle the limitation of convention field survey method for land-cover and yield estimation process. This research can be used to support the process of digital agriculture activities.
2. The convention land-use land-cover analysis and yield estimation were time-consuming and prone to biased conclusion. Application of satellite image data provide detail information and data processing can be done almost real-time. This approach would have significant contributions manage crop health, land cover land use analysis, early intervention and strategic planning.
3. On the other hand, semi-automatic training sample labeling causes poor classification performance due to mixed pixel values and small size training data. These problems have been handled by automated feature extraction techniques to discriminate one object from the other.
4. Advanced ML Algorithms like RF and SVM has been implemented with improved design of optimization function, the distance metric like Mahalanobis distance has been used in place of Euclidean distance to improve the efficiency, and similarly kernel functions are used to improve the performance of the linear SVM classifier.

Therefore, in this study a lot of attempts has been made to process raster image data, implementation of the satellite image datasets, selection and implementation of machine learning models to optimize classification performance. So, it is possible to conclude that, the proposed machine learning models were efficient to handle the limitation of conventional image classification system.

Recommendation

From the experiment result we obtained; the following recommendation has been drawn for further exploration in the domain area.

1. Less study has been reported in the case of Ethiopia's digital agriculture system; the domain area needs further research to improve and implement precision agriculture. In addition, the issues of high-resolution satellite image datasets need a solution for in-depth study.
2. The sensors like near infrared image sensor, light sensors, color sensors and NIR drones were not available easily in the local market. University should have some modality to mitigate the challenges of satellite image data purchasing.
3. From the experiment results, we have seen that most of base classifiers are sensitive to data variance and bias to handle the complexity problems, we recommend interested researchers to implement deep learning and ensemble learning approaches in the future.

REFERENCE

[1] N. Sulaiman *et al.*, “applied sciences The Application of Hyperspectral Remote Sensing Imagery (HRSI) for Weed Detection Analysis in Rice Fields : A Review,” 2022.

[2] J. Som-ard, C. Atzberger, E. Izquierdo-verdiguier, F. Vuolo, and M. Immitzer, “Remote Sensing Applications in Sugarcane Cultivation : A Review,” pp. 1–46, 2021.

[3] U. B. Gewali, S. T. Monteiro, and E. Saber, “Machine learning based hyperspectral image analysis: A survey,” 2018, [Online]. Available: http://arxiv.org/abs/1802.08701.

[4] M. Kanning, I. Kühling, D. Trautz, and T. Jarmer, “High-resolution UAV-based hyperspectral imagery for LAI and chlorophyll estimations from wheat for yield prediction,” *Remote Sens.*, vol. 10, no. 12, pp. 1–17, 2018, doi: 10.3390/rs10122000.

[5] F. Arias, M. Zambrano, K. Broce, C. Medina, and H. Pacheco, “Hyperspectral imaging for rice cultivation : Applications , methods and challenges,” vol. 6, no. 1, pp. 273–307, 2021, doi: 10.3934/agrfood.2021018.

[6] G. Elayaroja and U. Sankari, “A Survey on Hyperspectral Image Classification using Adaptive Spatial-Spectral Feature Learning,” vol. 6, no. 9, pp. 1–8, 2019.

[7] M. Tech, “Hyperspectral Sensor Data Fusion At Decision Level Using Support Vector Machine,” Pp. 14–18, 2016.

[8] M. Z. Jhandir and T. Ahmed, “Classification of cotton and sugarcane plants on the basis of their spectral behavior CLASSIFICATION OF COTTON AND SUGARCANE PLANTS ON THE BASIS OF THEIR SPECTRAL BEHAVIOR,” no. August, 2011.

[9] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, “Advanced Spectral Classifiers for Hyperspectral Images: A review,” *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, 2017, doi: 10.1109/MGRS.2016.2616418.

[10] J. Xue and B. Su, “Significant Remote Sensing Vegetation Indices : A Review of Developments and Applications,” vol. 2017, 2017.

[11] R. Tao, S. Member, X. Zhao, S. Member, W. Li, and S. Member, “Hyperspectral Anomaly Detection by Fractional Fourier Entropy,” vol. 12, no. 12, pp. 4920–4929, 2019.

[12] F. Melgani and L. Bruzzone, “Classification of Hyperspectral Remote Sensing,” vol. 42, no. 8, pp. 1778–1790, 2004.

[13] X. Ceamanos *et al.*, “A classifier ensemble based on fusion of support vector machines for classifying hyperspectral data To cite this version : HAL Id : hal-00578897,” 2011, doi: 10.1080/19479832.2010.485935.

[14] X. Liu *et al.*, “Large-Scale High-Resolution Coastal Mangrove Forests Mapping Across West Africa With Machine Learning Ensemble and Satellite Big Data,” vol. 8, no. January, pp. 1–15, 2021, doi: 10.3389/feart.2020.560933.

[15] P. Gong and P. J. Howarth, “An Assessment of Some Factors Influencing Multispectral Land-Cover Classification,” vol. 56, no. 5, pp. 597–603, 1990.

[16] G. Camps-valls and L. Bruzzone, “Kernel-Based Methods for Hyperspectral Image Classification,” vol. 43, no. 6, pp. 1351–1362, 2005.

[17] E. Adam, H. Deng, J. Odindi, E. M. Abdel-Rahman, and O. Mutanga, “Detecting the early stage of phaeosphaeria leaf spot infestations in maize crop using in situ hyperspectral data and guided regularized random forest algorithm,” *J. Spectrosc.*, vol. 2017, 2017, doi: 10.1155/2017/6961387.

[18] T. N. Pham, L. Van Tran, and S. V. T. Dao, “Early Disease Classification of Mango Leaves Using Feed-Forward Neural Network and Hybrid Metaheuristic Feature Selection,” *IEEE Access*, vol. 8, pp. 189960–189973, 2020, doi: 10.1109/access.2020.3031914.

[19] DADO, “Wheat Seed Production Techniques Manual,” 2016.

[20] J. Li, J. M. Bioucas-dias, A. Plaza, and S. Member, “Spectral – Spatial Classification of Hyperspectral Data Using Loopy Belief Propagation and Active Learning,” vol. 51, no. 2, pp. 844–856, 2013.

[21] S. M. Ganie, M. B. Malik, and T. Arif, “Various Platforms and Machine Learning Techniques for Big Data Analytics : A Technological Survey,” vol. 3, no. 6, pp. 679–687, 2018.

[22] H. N. Trong, T. D. Nguyen, and M. Kappas, “Land Cover and Forest Type Classification by Values of Vegetation Indices and Forest Structure of Tropical Lowland Forests in Central Vietnam,” vol. 2020, 2020.

[23] Q. Zheng, W. Huang, X. Cui, Y. Shi, and L. Liu, “New spectral index for detecting wheat yellow rust using sentinel-2 multispectral imagery,” *Sensors (Switzerland)*, vol. 18, no. 3, pp. 1–19, 2018, doi: 10.3390/s18030868.

[24] T. S. Rathna Priya and A. Manickavasagan, “Characterising corn grain using infrared imaging and spectroscopic techniques: a review,” *J. Food Meas. Charact.*, vol. 15, no. 4, pp. 3234–3249, 2021, doi: 10.1007/s11694-021-00898-7.

[25] G. Truth, “Gaussian Processes for Vegetation Parameter Estimation from Hyperspectral Data with Limited Ground Truth,” 2019, doi: 10.3390/rs11131614.

Part I

A Questionnaire to collect field data for actual yield of the selected study area

Respondent Type: **Farmer** **Expert** **Agent**

1 What is the area of selected agricultural land in square in square meters?
2 What are the major land cover in the study area and how land cover analysis conduct by the domain expert.

3 Which crop(s) is/are commonly sown in the selected agricultural land?

Crop A ……
Crop B ……
Crop C …….

4 What is the month of sowing of the identified crops in the selected agricultural area?

Crop A ……
Crop B ……
Crop C …….

5 What are the months or period critical for harvesting crops in the study area?
_____
6 What is total Labor cost per meter at the time sowing and harvesting

Answer ……..

7 How often you utilized fertilizer, types of fertilizer suitable and method of utilization.
_____

8 Expected/Estimated Number of grain per square meter

Crop A ……
Crop B ……
Crop C …….

9 Actual Number of grain roots per square meter……
Crop A ……
Crop B ……
Crop C …….
10 Number of seeds per grain……

Crop A ……
Crop B ……
Crop C ……

11 What are the techniques employed to compute a yield?

_____

12 Total yield per square meter……

Crop A ……
Crop B ……
Crop C ……

13 What are common crop diseases which affects your crop production and how you measure the severity level?

_____

14 Adverse factors for the yield loss and its impact on food security?

_____

Part II

Bands DN to reflectance model

Band	RADIANCE_MULT-BAND	RADIANCE_ ADD-BAND	REFLECTANCE_ MULT-BAND	REFLECTANCE_ ADD-BAND
1	1.2983E-02	-64.91656	2.0000E-05	-0.100000
2	1.3295E-02	-66.47535	2.0000E-05	-0.100000
3	1.2251E-02	-61.25646	2.0000E-05	-0.100000
4	1.0331E-02	-51.65490	2.0000E-05	-0.100000
5	6.3220E-03	-31.61022	2.0000E-05	-0.100000
6	1.5722E-03	-7.86118	2.0000E-05	-0.100000
7	5.2993E-04	-2.64964	2.0000E-05	-0.100000
8	1.1692E-02	-58.45913	2.0000E-05	-0.100000
9	2.4708E-03	-12.35399	2.0000E-05	-0.100000
10	3.3420E-04	0.10000	**----**	**----**
11	3.3420E-04	0.10000	**----**	**----**