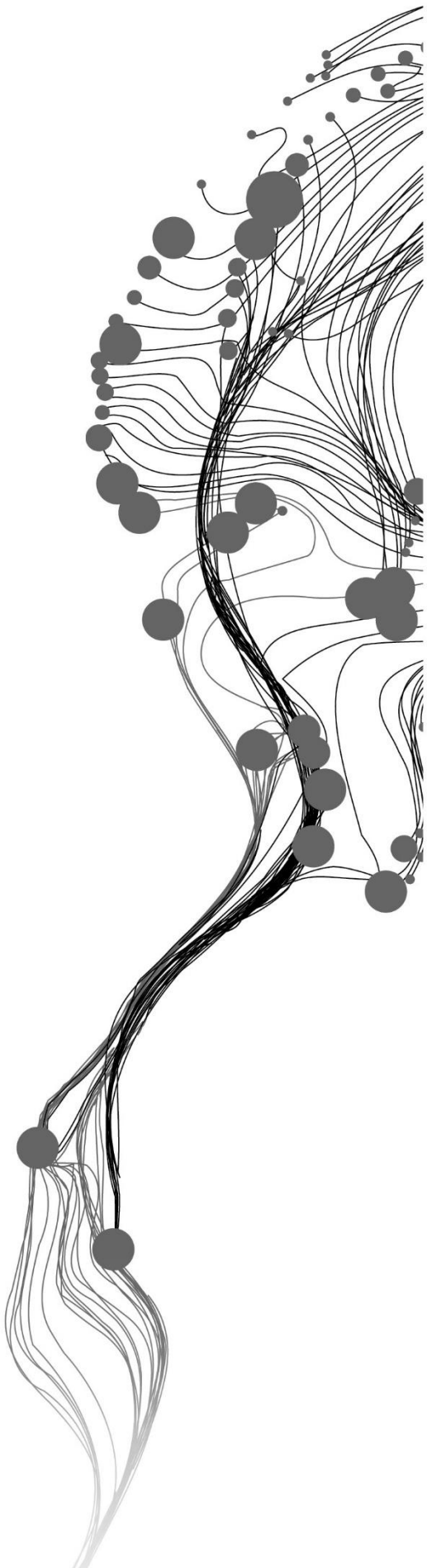
Analysis of Machine Learning Classifiers for LULC Classification on Google Earth Engine

SHOBITHA SHETTY

March, 2019

SUPERVISORS:
Mr. Prasun Kumar Gupta
Dr. Mariana Belgiu
Dr. S.K.Srivastav



Analysis of Machine Learning Classifiers for LULC Classification on Google Earth Engine

SHOBITHA SHETTY
Enschede, The Netherlands, March, 2019

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Mr. Prasun Kumar Gupta
Dr. Mariana Belgiu
Dr. S.K.Srivastav

THESIS ASSESSMENT BOARD:
prof.dr.ir. A.Stein (Chair)
Mr. Pankaj Bodani (External Examiner, SAC, Ahmedabad)

ABSTRACT

Classifiers that provide highly accurate Land Use Land Cover (LULC) maps are always in demand when reliable information is required from remotely sensed images. Machine Learning Classifiers in particular produce good classification results even on high-dimensional, complex data. Accuracy of the classified maps are affected by various factors such as training sample size, quality of training samples, thematic accuracy, choice of classifier, study area size and so on. Understanding these factors will help in obtaining best possible classification accuracies for a given requirement. Classification tasks involving large number of satellite images and features become a computation intensive process leading to Big Data problems. Recently, free cloud-based platforms such as Google Earth Engine (GEE) provided parallel-processing environments to perform any such tasks related to image classifications. The current research mainly uses GEE to analyse various machine learning classifiers such as Classification and Regression Trees (CART), Random Forest, Support Vector Machine (SVM), Relevant Vector Machine (RVM) using multi-temporal Landsat-8 images and compares their performance under the influence of data-dimension, sample size and quality. GEE's support to external programs facilitated the integration of an unavailable GEE classifier, RVM. RVM, a potential sparse Bayesian Classifier is reported to perform similar to SVM from previous works. With its probabilistic output, RVM performed comparatively well with as small sample size as 2 pixels/class with an overall accuracy of 64.01%. Though RVM has been evaluated before with SVM, studies on relative performance of RVM with other mature classifiers such as Random Forest and CART was missing. The study found that RVM was comparable to SVM at very small sample sizes, but CART and Random Forest performed better than RVM by 10-20%. While tree-based CART and Random Forest classifiers performed well under the influence of various factors, kernel-based SVM and RVM classifier performed well with smaller training samples. These classifier performances are also highly affected by the quality of training samples. With less available studies focusing on training samples and practical limitations of ground data collection for large study areas, study of sampling techniques has advanced as an important topic. Three different stratified sampling techniques were analysed and Stratified Equal Random Sampling performed better than Stratified Proportional Random Sampling method for smaller classes by reliably mapping even the rare classes. Though the imbalanced dataset created by Proportional Random Sampling gives a better overall accuracy, the former method performs well in all cases. If the aim is to obtain uniform spread of samples by considering the underlying variations of a class, Stratified Systematic sampling can be used. This method utilises semi-variance and Spatial Simulated Annealing process to obtain an optimal sampling scheme over a region resulting in good class-level accuracies. The only disadvantage of the method is the probability of error propagation with erroneous first sample. Overall, though the study understands the choice of classifiers depend on the requirement, Random Forest can be considered as a universal classifier with more than 95% confidence since it outperformed all 3 algorithms in all scenarios.

Keywords: *LULC, Random Forest, Support Vector Machine, CART, GEE, Classification, Machine Learning, Sampling, Accuracy, Random, Systematic Sampling, Spatial Simulated Annealing*

ACKNOWLEDGEMENTS

Throughout the journey of the MSc research, the neural network in me performed a good amount of learning based on the various observations and analysis. Though I cannot measure the accuracy of my LULC here, I am very grateful to lot of people for helping and guiding me throughout this stage.

Firstly, I am highly indebted to my supervisors *Mr. Prasun Kumar Gupta*, *Dr. Mariana Belgiu* and *Dr. S.K. Srivastav* for their timely guidance and support. Mr. Prasun Kumar Gupta has been a constant support throughout the research period, guiding and patiently moulding me towards it. He has always encouraged new ideas and instilled a confidence in me by his calm nature. Dr. Mariana Belgiu had her door always open for any issues and is instrumental in helping me develop the research skills. With an open mind she would listen to any of my discussions and steer me in the right direction when required. Dr. S.K. Srivastav would always take out time from his very busy schedule and provide valuable suggestions on my research work. He helped me understand how to execute an organised research which was vital for completion of the work. I thank all my supervisors for pushing me towards the right direction and encouraging me to never give up. It was an honour for me to be your student.

I would like to extend my heartfelt gratitude to *Dr. Sameer Saran* who always had our back during the whole MSc course, making sure that everything is in place to ensure our smooth progress. I would like to thank all the faculties of IIRS and ITC for showering us with your wide knowledge and helping us learn. Well, a big thanks goes without saying to all my MSc-PGD *batch mates* who would happily share their precious time for any discussions and support each other during the entire process.

I am greatly thankful to the Google Earth Engine team and developers group for their inputs. I would like to particularly thank *Noel Gorelick*, for patiently answering my queries.

This acknowledgment will not be complete without extending my gratitude to my family and friends. My sincere thanks to my parents, brother and sister-in-law who supported me on my decisions and encouraged me throughout the MSc course. I would like to particularly mention *Amma*, for being my backbone throughout, helping me stay strong and keep going. My friends, *Aparna* and Anirudha, for standing by me always and making me believe in myself.

- Shobitha Shetty

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Motivation and Problem Statement

Land use land cover (LULC) explains the various land features present on the surface of the earth. While land use gives an indication of how the land is being used for different purposes such as agriculture, industrial areas and residential areas, land cover refers to the physical land types such as settlements and built-up areas, forests, water bodies and grasslands. Understanding of LULC at regional and global scales leads to the study of various processes that affect the earth such as flood, climate change, erosion, and migration. Directly or indirectly, land use land cover is an indicator of underlying trends of different natural and social phenomena (Lam, 2008).

From climatic changes to policy planning, LULC information plays a critical role in their understanding. While government institutions use LULC for planning policies (Johnson, Truax, & O'Hara, 2002), subsidies, urban planning and development (Thunig et al., 2011), other agencies use LULC for several applications such as monitoring crop yield and productivity (Lobell, Thau, Seifert, Engle, & Little, 2015), monitoring of forest degradation and illegal logging activities (Hansen et al., 2013), management of critical animal habitats, biological hotspots and restoration and rehabilitation under disaster management (Guru & Aravind, 2015). Studying LULC changes gives crucial information regarding the current global issues, such as melting of ice, changes in rainfall patterns, abnormal temperatures, urban sprawl (Aboelnour & Engel, 2018), conflicts and food security (Abdulkareem, Sulaiman, Pradhan, & Jamil, 2018). There are many other disciplines in which LULC information plays an indispensable role and extracting LULC data has been an important and ongoing field of research since many years.

The most feasible solution for extracting LULC information is by classifying the images obtained from remote sensing (Shalaby & Tateishi, 2007). Image classification involves assigning pixels to a particular class based on various features, such as spectral signatures, indices, contextual information etc. Many classification techniques exist among which Maximum Likelihood Classification (MLC) was the most popular parametric classifier due to its excellent classification results (Yu et al., 2014). But the parametric classifiers assume normal distribution of data and in real-world data do not follow such distribution. In recent decade, machine learning classifiers have emerged as powerful classifiers and have been widely adopted for LULC classification due to their higher accuracy and performance compared to MLC (Ghimire, Rogan, Galiano, Panday, & Neeti, 2012). These non-parametric classifiers do not have any assumptions about the data distribution.

Non-Parameteric Machine Learning classifiers such as Random Forest (RF), Support Vector Machine (SVM), Classification and Regression Trees (CART) have been reported to provide highly accurate LULC classification results using remotely sensed images (Foody & Mathur, 2004; Nery et al., 2016). CART is a simple binary decision tree classifier used in the classification of few global LULC maps. It works by recursively splitting the node until terminal nodes are reached according to a pre-defined threshold (Tso & Mather, 2009). Though CART tends to over fit the model to some extent, it's fast performance and accurate results make CART as one of the widely used LULC classifiers (Lawrence & Wrlght, 2001). RF is an

ensemble classifier formed by the combination of multiple CART-like trees. Each tree independently classifies the data and votes for the most popular class (Breiman, 2001). Additionally, RF follows bagging approach where each tree samples a subset of feature and training data with replacement. Its light computation and highly accurate results make RF one of the favourite classifier for LULC (Gislason, Benediktsson, & Sveinsson, 2006). SVM is another well performing classifier which tries to build an optimal hyperplane separating various classes with minimum misclassified pixels in training step. For non-linear datasets, SVM projects the data into another higher dimensional feature space using the kernel trick. The accuracy is highly affected by the choice of kernels and other input parameters, which is one of the major disadvantages of SVM (Mountrakis, Im, & Ogole, 2011).Despite this, SVM is one of the popular classifier in remote sensing field that performs well by selecting a smaller subset of support vectors from training samples (Giles M. Foody, Mathur, Sanchez-Hernandez, & Boyd, 2006). Relevance Vector Machine (RVM) is another machine learning classifier developed by Tipping (2001) that follows Bayesian approach for learning the training sample patterns and provides probabilistic output about each class. RVM is a similar functional form of SVM but without the disadvantages of using complex mercer kernels and trade off due to parameter estimation . While SVM aims to maximise the marginal distance between classes and determine support vectors at the boundaries of a class, RVM attempts to find relevant vectors with non-zero posterior probability distribution of their weights (Tipping, 2001). Such relevant vectors are usually found away from the boundaries of a class. RVM is reported to perform better than SVM in terms of classification results with fewer training samples (Pal & Foody, 2012). However, studies on RVM in remote sensing are very scarce. Furthermore, there are no studies which compare RVM with other machine learning classifiers such as RF and CART.

Additionally, accuracies of such advanced methods are influenced by the choice of training samples, sample size and heterogeneity of samples. Stehman (2009) investigated on various sampling designs and ways for assessing accuracy of the existing classification. However, this study did not deal with the impact of sampling method on obtaining training data. Heydari & Mountrakis (2018) in their research mentioned the importance of understanding the sampling method in identifying accurate land cover classes. Giles M Foody (2009) describes how small change in sampling size brings a difference in the overall accuracy. Understanding the influence of the various factors such as training samples and their size, landscape heterogeneity helps reduce their negative influence on classification results and thereby greatly increase the accuracy.

With the free availability of high resolution and multi-temporal remotely sensed images from Sentinel-2 (Four 10m resolution bands with 5 days revisit) and Landsat 8 (30m resolution with 16 days revisit), LULC classifiers can be used to understand the issues related to flood, drought, urbanization agriculture and so on at national, continental, global levels (Pielke et al., 2011) and, thus to generate LULC products at better resolutions than currently available. For example, most of the global land cover products such as GlobCover have a coarser resolution such as 300m (Bicheron et al., 2008). Implementing and analysing the relation between the various factors and machine learning classifiers on higher resolution and multi-temporal images at larger scale is not a trivial task as it requires powerful machines which can manage high volumes of data and complex computations in considerable amount of time. Such high computing environments are usually only accessible to certain audience . Additionally, this prevents frequent updating of LULC maps at larger scales. While state-of-the art classifiers such as RF require high computation with increasing number of trees and features, SVM becomes more processing intensive on larger datasets and so does RVM. To help with the Big Data and computation problems in recent years, few cloud platforms are made available which provide required environments for processing large data along with a repository of images for LULC classification. Particularly Google Earth Engine (GEE) is emerging as a powerful tool which has been used

in few studies on classification for specific LULC applications related to crop and urban at regional and global scales (Dong et al., 2016). GEE also provides a way for collaborating and working on the same pre-processed data set, thus paving a way for different users to re-use or validate the concepts.

This project focuses mainly on identifying most suitable machine learning classifier for large scale LULC maps in terms of accuracy and sampling designs for training data, on a cloud based platform.

1.2. Research Idenfication

LULC data at larger scales which has sufficient, updated and accurate details is of utmost importance. Higher classifier accuracy can be obtained by considering a series of composite satellite images which provides data with minimal noise. Nevertheless, generating and processing such data requires high-computing environment (Lück & van Niekerk, 2016). This research intends to use these data to implement and compare the existing machine learning image classifiers on a cloud platform like GEE using Landsat 8 multi-temporal images, along with understanding the ways of sampling training data and its effectiveness. The best suited LULC classifier and training sampling method can then be scaled and applied on a higher resolution image covering larger regions. The outcome of this research can be used by other researchers to perform LULC analysis on large scale data.

1.2.1. Research Objectives

The main aim of this research is to understand the two main aspects of LULC classification – Training sample selection and analysis of machine learning classifiers. Thus the research aims to attain the following objectives:

1. Understand the effect of different stratified training sampling methods using machine learning classifiers
2. Integrate and analyse RVM algorithm on the GEE cloud platform
3. Assess the efficiency of various machine learning classifiers which satisfy the thematic class definition provided by International Geosphere Biosphere Program (IGBP), on GEE
    a. The performance of Random Forest, SVM, CART to generate LULC maps from multi-temporal images will also be evaluated in this study.
    b. Compare the accuracies of RVM with evaluated machine learning classifiers

1.2.2. Research Questions

The above research objectives can be reached by defining a solution to the following research questions:

1. What is the effect of different training sampling techniques on the accuracy of LULC classification?
    a. Do the sampling methods equally affect smaller and larger sized classes?
    b. What are the advantages and disadvantages of different methods?
2. How does RVM perform in classifying different LULC classes and What is the effect of training sample size on the classification result?
3. How well do the in-built machine learning methods of GEE such as Random Forest, SVM and CART perform on multi-temporal satellite images in discriminating land cover classes of interest?
    a. Which is the overall best performing classifier
    b. How well do the classifiers perform with respect to each other?

c. To what extent does the integrated RVM classifier perform compared to RF, CART and SVM?

1.2.3. Innovation Aimed At

By concentrating on machine learning classifiers for LULC, the project brings the following novelty to the research.

1. Exploring a potential machine learning classification technique, such as RVM by integrating it in a cloud based platform to study its accuracies for LULC will be first of its kind
2. Comparative performance of RVM with other machine learning classifiers such as CART, RF which has not been analysed before
3. Studying the effect of choosing training samples using stratified systematic sampling method

1.2.4. Research Approach

The general research methodology is shown in Figure 1-1. Most of the research processes are performed in GEE. The Landsat 8 dataset available in GEE public data catalogue is imported and images are pre-processed to capture various multi-temporal data. Once the dataset is ready, three different sampling techniques are applied to obtain training samples. The training samples are used for training four different machine learning classifiers i.e. RF, CART, SVM and RVM. Among them, RVM classifier is integrated to GEE using a python implementation. Finally, different accuracy assessment methods are followed to evaluate the sampling techniques and machine learning classifiers. A detailed methodology flow diagram is shown in Figure *4-1* of Chapter 4.

1.3. Thesis Outline

The thesis is organised into six chapters. **Chapter 1** describes the motivation behind the research and the problems being addressed. It also briefs about the research objectives, research questions and the innovation in the project. **Chapter 2** gives information from previous literatures regarding LULC, machine learning classification techniques, sampling methods and few more. **Chapter 3** discusses about the study area and datasets. **Chapter 4** describes in detail the different methods followed to achieve the research objectives. While **Chapter 5** depicts the outcome of these methods, **Chapter 6** discusses further about the results. **Chapter 7** concludes by answering the research questions and future recommendations.

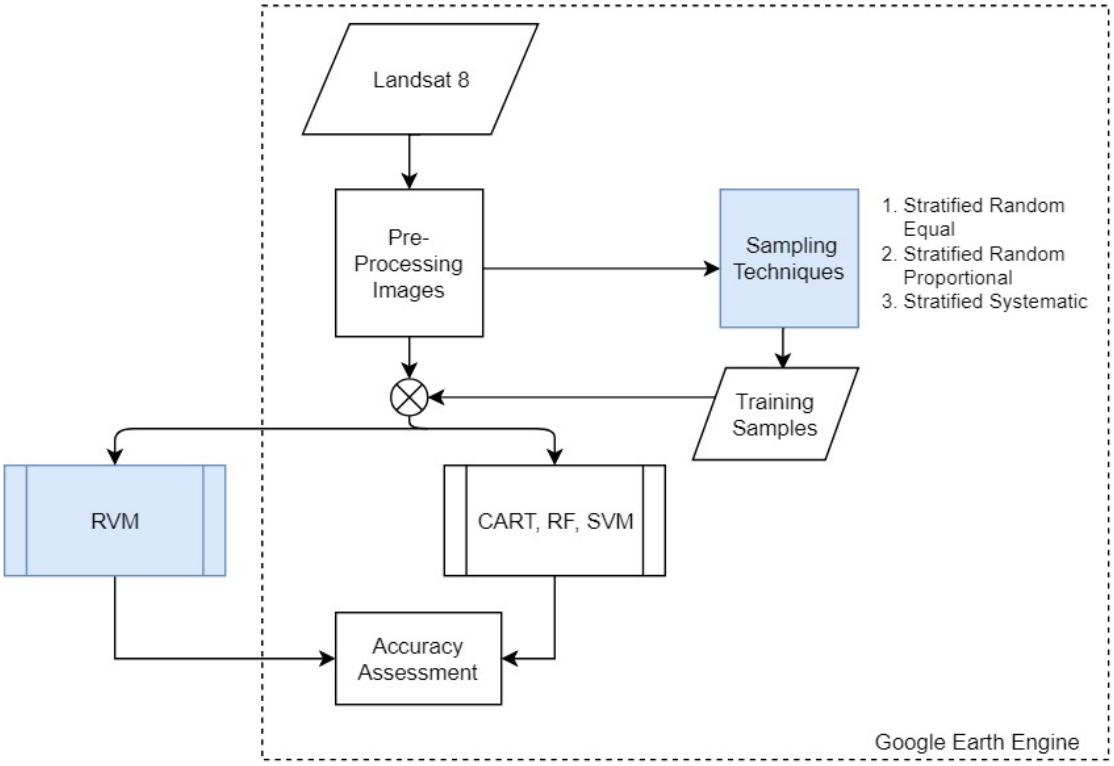


Figure 1-1:General Methodology

2. LITERATURE REVIEW

2.1. Land Use Land Cover Classifiers

Extracting accurate LULC data from remotely sensed images require good image classification techniques. In general, these classifiers can be grouped as supervised and unsupervised, or parametric and non-parametric, or hard and soft (fuzzy) classification, or per-pixel and subpixel based classifiers. Many classifiers exist whose performance are affected by various factors such as choice of training samples, heterogeneity of study area, sensors, number of classes to identify and so on. (Lu & Weng, 2007). Creation of more accurate maps is always a necessity. As a result, new classification methods keep adding to the list in various literatures. A systematic comparative analysis of different algorithms is important to identify the improvement in the new classifiers. Certain literatures (Khatami, Mountrakis, & Stehman, 2016; Lu & Weng, 2007) and books (Tso & Mather, 2009) provide comprehensive information about different classifiers. Among all the classifiers, Yu et al. (2014) found in his studies that the parametric MLC is most often used for image classifications, Though in recent decades machine learning classifiers have been reported to perform well,

2.1.1. Machine Learning Classifiers

Machine Learning is among the most reliable approaches for classification of non-linear systems. It helps understand the behaviour of a system based on the input observations and has the ability to approximate the values without the prior knowledge of the relationship between the data. This makes machine learning technique a suitable choice in classification of remote sensing images where it is impossible to have a complete knowledge of the characteristics of the whole study area (Walker, 2016) . Thereby, with the advent of complex data and easy availability of higher resolution satellite imageries, machine learning classifiers are already increasingly used in the remote sensing field (Pal & Mather, 2004; Pal & Mather, 2005).

Machine Learning Classifiers are reported to produce higher accuracy even with complex data and higher number input features. (Aksoy, Koperski, Tusk, Marchisio, & Tilton, 2005; Huang, Zhou, Ding, & Zhang, 2012). Few of the popular classifiers are CART, RF, k-Nearest Neighbor (k-NN), SVM, Artificial Neural Network (ANN) etc. While some of these classifiers such as CART build simple decision tree from the given training data, RF uses random subset of training data to construct multiple decision trees. Other classifiers such as ANN follow a neural network pattern and build multiple layer of nodes to passes input observations back and forth during the learning process (Multi-Layer Perceptron) until it reaches a termination condition (Mas & Flores, 2008). k-NN's use the information about the neighboring pixels to develop an understanding of the underlying pattern of the training dataset (Calvo-Zaragoza, Valero-Mas, & Rico-Juan, 2015).

On the other hand classifiers such as SVM find a subset of training data as support vectors by fitting a hyperplane that separates twos classes in the best possible way (C. Huang, Davis, & Townshend, 2002). Among all these classifiers, most literatures suggest that RF and SVM have an upper hand in most classification scenarios as they outperform other machine classifiers (Belgiu & Drăguţ, 2016; Nery et al., 2016). However, there has been a lesser known machine learning classifier published by Tipping (2001), RVM, which has been reported to perform better than SVM from the few available studies (Mountrakis, Im, & Ogole, 2011; Pal & Foody, 2012). Hence there is a need to explore RVM further in order to understand its performance for LULC classification and in-comparison with other machine learning classifiers. The next sub-section describes in detail the various machine learning classifiers, which are the subject of the current study.

2.1.1.1. Classification and Regression Trees

CART is among the simplest binary classifier developed by Breiman, Friedman, Olshen, & Stone (1984) which works based on the framework of hierarchical decision trees. The main advantage of such structures is that classification decisions can be treated as a white box system, where the input-output relations can be understood and interpreted easily compared to multilayer neural networks (Tso & Mather, 2009)

The input and output of the CART algorithms are connected by a series of nodes, where each node is split into two branches, finally leading to leaf nodes that represent class labels in case of classification trees, and continuous variables in case of regression trees. The repeated split of nodes proceeds until it reaches a threshold criterion. CART uses Gini Impurity Index to decide the input features which will provide the best split at each node (Tso & Mather, 2009). The split can be univariate, where decision boundaries are parallel to input feature axis, or multivariate, which is a linear combination of input features (Tsoi & Pearson, 1991). Multivariate decision boundaries provide more flexibility to each class boundary.

CART tends to over-fit the tree when it particularly fits the training data better. This is overcome by pruning the tree so that it can be robust to the non-training input data. CART uses cross-validation technique for pruning which reduces those branches whose removal does not affect the results beyond a defined threshold (Lawrence & Wrlght, 2001). This might lead to decrease in accuracy for classification of training data and loss of certain information, but on the other hand results in increase of accuracy for unknown data (Pal & Mather, 2003)

Tree-based classifier such as CART are widely used in various studies in the remote sensing field, e.g., MODIS global land cover data is developed using CART due to its robustness and simplicity (Friedl et al., 2002). Certain other studies used CART and SVM for natural habitat mapping with similar performance from both the algorithms (Boyd, Sanchez-Hernandez, & Foody, 2006). Bittencourt & Clarke (2003) work on CART showed good results for spectrally small and similar AVIRIS dataset. Lawrence & Wrlght (2001) through their studies indicate that CART has a major advantage of automatically choosing those input and ancillary data that are useful for classification. Additionally, CART provides probability of misclassification at every leaf node, thus helping in assessing the quality of assessment. For lower dimensional data, CART performs faster than Neural Networks and gives comparable results (Pal & Mather, 2003). On the other hand, CART is highly sensitive to sample size chosen for each class. High dimensionality data also reduces the performance of CART as it leads to complex tree structures.

2.1.1.2. Random Forest

Tumer & Ghosh (1996) proved that combining output from multiple classifiers for predicting an outcome gives very high classification accuracies. This is the basis for the ensemble classifier RF, which combines output from multiple decision trees to decide the label for a new input data based on maximum vote.

Random Forest randomly selects a subset of training sample through replacement to build a single tree, i.e it uses bagging technique where for every tree, data is sampled from original complete training set. This might result in same samples being selected for different trees while others not being selected at all (Breiman, 1996). The samples that are not used for training (out-of-bag samples) are internally used for evaluating the performance of the classifier and provides an unbiased estimate of the generalization error. Furthermore, at each node RF performs the random selection of variables from training samples to determine the best split to construct a tree. Though this can decrease the strength of individual trees, it reduces the correlation between the trees resulting in lower generalization error (Breiman, 2001). To choose the best split, RF uses

Gini Index measure which gives a measure of impurity within a node. The split is performed in such a way that there is a decrease in entropy and increase in information gain after the split. But the performance of tree based classifiers are more affected by the choice of pruning methods than the best split selection measure (Pal & Mather, 2003). RF is immune to such affects as it builds trees without the need to employ pruning techniques (M Pal, 2005).

One of the user defined parameters for RF is the number of trees. Breiman (1999) suggests that the generalization error always converges as the number of trees increase. Hence there is no issue of overfitting which can also be attributed to Strong Law of Large Numbers (Feller, 1971). Thus for RF, number of trees can be as large as possible but beyond a certain point, additional trees will not help in improving the performance of the classifier (Guan et al., 2013). Belgiu & Drăguț (2016) suggest in their review that most papers use 500 number of trees for RF classification while there are few other studies which use 5000,1000 or 100 trees for RF. And among these, 500 is considered as the acceptable optimal value for number of trees. Number of variables required to decide the best split is another user-defined parameter which highly affects the performance of RF. And this is usually set to square root of the number of input variables.

A single tree may not capture the importance of all the input features and might favour certain features during classification but a combination of trees takes into account all the features which are randomly selected from the training samples. Thus, in terms of remote sensing RF helps in understanding the relative importance of different variables derived from the bands of a satellite image . RF assesses each variable by removing one from randomly chosen input variables while keeping other variables constant. It estimates the accuracy based on out-of-bag error and Gini Index decrease (Ghosh, Sharma, & Joshi, 2014). Additionally, RF also measures the proximity of two samples chosen based on the number of times the pair ends up in the same terminal node. This proximity analysis helps in detecting incorrectly labelled training samples and makes RF insensitive to noise (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012).

RF has gained its importance due to its robustness to noise and outliers. Furthermore, RF performs better than other classifiers which use ensemble methods such as bagging and boosting (Gislason et al., 2006). RF has even proven to give good results when used in various applications such as urban landscape classification (Ghosh et al., 2014), land cover classification on multi-temporal and multi-frequency SAR data (Waske & Braun, 2009) and so on.

2.1.1.3. Support Vector Machine

SVM is one of the widely used classifiers in remote sensing field. SVM gained its importance due to highly accurate classification results with lesser training samples, which is usually a limitation in land use land cover classification scenarios (Mantero, Moser, Member, Serpico, & Member, 2005).

SVM is a linear binary classifier which is based on the concept that training samples which are at closer proximity to the boundaries of a class will discriminate a class better than other training samples. Hence SVM focuses on finding an optimal hyperplane which separates the input training samples of various classes. The samples present close to the boundaries of a class and at minimum distance to the hyperplane are taken as support vectors, which are used for the actual training. Figure *2-1* shows a case where classes are linearly separable and hence the support vectors lie on the decision boundary. But this is not usually the case. For classes that share a non-linear relationship, a relaxation is introduced in the form of slack variable $\xi \geq 0$ which allows few incorrect pixels within a class boundary while achieving a hyperplane. Furthermore, to balance the trade-off between misclassification errors and the margin, there is a user-defined cost parameter

C which controls the penalty applied on misclassified pixels. This results in the creation of soft margin hyperplanes (Cortes, Vapnik, & Saitta, 1995).

Cost Parameter C, highly influences the selection of support vectors and performance of SVM. Few literatures suggest exponentially changing the C value using a grid search method to find an optimal C.
Low value of C allows for more misclassified pixels to be present in a class and tends to include more support vectors, which can lead to lower classification accuracies. While very high values of C will result in overfitting and generalization error (Foody & Mathur, 2004).

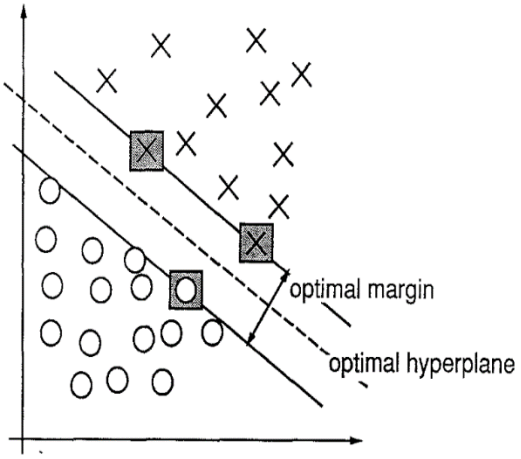


Figure 2-1: Support Vectors and the hyperplane in 2-Dimensional Space. Source Adapted from: (Cortes, Vapnik, & Saitta, 1995)

Another technique adapted to deal with non-linear input data(x) is the transformation of an input space to another higher dimensional feature space where, the training samples can be linearly separated. This transformation is achieved through a kernel trick where a mapping function Φ transforms x into Φ(x). (Boser, Guyon, & Vapnik, 1992). Training problem appears in the form of dot product of two vectors ($\Phi(x_i)$, $\Phi(x_j)$). The computational cost of higher dimensional space ($\Phi(x_i)$, $\Phi(x_j)$) is less because the following kernel transformation k is applied as shown in equation 2.1.

$$\Phi(x_i), \Phi(x_j)) = k(x_i, x_j) \quad 2.1$$

Additionally, this has added advantage that the knowledge of the mapping function is not needed (Huang, Davis, & Townshend, 2002). Only the user has to choose a kernel which follows Mercer's Theorem. Various kernel functions exist such as polynomial kernel, linear kernel and radial basis kernel (RBF). The choice of kernels also affect the results of the classification. Kernels such as RBF has a user-defined γ parameter which controls the influence of a training sample on the decision boundary. Higher the value of γ, more tightly fit are the decision boundaries around the samples. But this can lead to overfitting. Hence it is necessary to strike a right balance (Foody & Mathur, 2004).

The influence of user-defined parameters is also discussed by Mountrakis et al. (2011) in their review of support vector machines where they conclude the choice of kernels being a major setback of SVM. This is evidenced by the different results obtained from different kernels. Furthermore, the choice of C and γ highly influences the output. While there are certain literatures which suggest ways of handling the kernel issues (Marconcini, Camps-Valls, & Bruzzone, 2009), studies are very scarce which describe a standard way to choose such parameters (e.g., Chapelle & Bousquet, 2002). However, SVM, a non-parametric classifier, is still among the popular classifiers as it gives highly accurate results with limited training samples while generalizing well on new input data. It also works well with higher dimensional data which is a good advantage in remote sensing field as more and more higher resolution, multi-spectral data are made available (Srivastava, Han, Rico-Ramirez, Bray, & Islam, 2012).

SVM is also widely used to solve multi-class classification problems using one-against-all and one-against-one techniques. While one-against-all compares one class with all other classes taken together, generating *n*(number of classes) classifiers, one-against-one forms $(n(n-1))/2$ classifiers by forming all two-class classifier pairs from the given input classes (Pal & Mather, 2005; Xin Huang & Liangpei Zhang, 2010)

2.1.1.4. Relevant Vector Machine

RVM is a Bayesian form of linear model and probabilistic extension of SVM developed by Tipping (2001) which provides sparse solution to classification tasks. Bayesian inference approach treats feature set 'w' (weights) related to input observations as random variables and understands the distribution of these weights w.r.t given input and target data. This posterior probability distribution of w helps predict the target values for any new input data. The most useful part of this Bayesian approach is that it removes all the irrelevant variables and creates a simple model which explains the pattern in the data (Tipping, 2004). This is an attractive feature for classification in remote sensing where it is difficult to get abundant training samples.

The study by Tipping (2001) explains the RVM process as summarized in the following section. For a given training set $\{x_n;t_n\}$ where x_n and t_n are input and target values respectively, RVM concentrates on finding probabilistic distribution of values of w in the model shown in equation 2.2 such that *y(x)* generalises any new input observation. Here y(x) represents a function defined over input space (target values), Φm represents the basis functions, M represents the number of variables and w_m represents the set of variables associated with observations and € represents Gaussian-noise with variance σ^2.

$$y(x) = \sum_{m=1}^{M} w_m \varphi_m \ (x) + \epsilon \qquad 2.2$$

The y(x) problem reduces to estimating the conditional probability distribution of target values based on the parameter distribution. Using already known mapped values of input and targets, the posterior probability distribution of parameter w can be found. To control overfitting the model, a Gaussian prior α is defined over w and independent gamma hyper prior is defined over α and variance. RVM is a binary classifier that uses Bernouilli distribution to find the value of w that maximizes the probability of finding good results. with a logistic sigmoid function as show in equation 2.3 and 2.4. Further, equation 2.3 is an extension for multi-class classification of RVM where k represents the number of classes. Since the maximum likelihood estimation is computation intensive, Tipping & Faul (2003) introduced a faster version of the same by controlling the way basis functions are deleted from the model.

$$\mathrm{p}\,(\mathrm{y} \mid w) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \prod_{\mathrm{j}=1}^{k} \sigma\left\{\left(\mathrm{y_j}(\mathrm{x_i})\right)\right\}^{\mathrm{y_{ij}}} \qquad 2.3$$

$$\sigma\big(\varphi(\mathrm{x})\big) = \frac{1}{1 + exp(-\varphi(\mathrm{x}))} \qquad 2.4$$

$$\mathrm{f(w)} = \sum_{i=1}^{n} \log \mathrm{p}(y_i(w_\mathrm{i})) \ + \sum_{i=1}^{n} \log \mathrm{p}(w_i(\alpha_\mathrm{i})) \qquad 2.5$$

During the training process, the priors act as penalty terms on the input observations and iterative analysis is performed to find p(y|w). If α_i represents maximum a posteriori (MAP) estimate for hyperparameter, MAP for weights are obtained by maximizing the equation 2.5 representing likelihood of class labels and prior on weights. As a result, most of weights get associated with very large values of α makes the corresponding vectors irrelevant. This results in creation of sparse model which considers only relevant vectors (non-zero co-efficient of w) to further estimate the probability distribution of weight.

RVM started as binary classifier and just like SVM it can extended for multi-class classification using one-against-all strategy. RVM gives similar accuracy results as SVM during image classification(Pal & Foody, 2012). Like SVM, RVM performs well with smaller training samples. Though highly popular, SVM has many disadvantages which is overcome by RVM and discussed in studies such as Foody (2008), Mountrakis et al. (2011), Pal & Foody (2012) :

- SVM uses more basis function than necessary. This makes it computationally complex. RVM on the other hand RVM uses much less vectors than SVM making it a sparser approach than RVM.
- SVM gives hard output as classification result while RVM gives a probabilistic output. This helps to analyse the uncertainty of each class.
- SVM is highly sensitive to user defined cost parameter C. But the parameters are automatically estimated in RVM.
- SVM requires Mercer's kernels which are complex and computation intensive but RVM works based on simple non-Mercer kernels and still gives similar accuracies.

2.1.2. LULC Classifiers Comparisons

With the increase in demand for accurate LULC data from remotely sensed images, it is important to understand the performance of various machine learning classifiers with respect to each other. Most of the studies performed a comparative analysis by focusing on the classifier. For example, studies such as Gislason et al., (2006), Mochizuki & Murakami (2012) concentrate on evaluating the performance of different tree based classifiers such as RF, CART, other decision trees with bagging/boosting approach (AdaBoost), wherein RF has outperformed other classifiers. Though CART tends to over-fit a model, in terms of training speed the simple binary tree structure of CART makes it faster than other machine classifier such as ANN and SVM (C. Huang et al., 2002). Lu & Weng (2007) made an in-detail study of all factors that are generally related to image classification techniques and found that the success of image classification depends on sources of data, effect of scale and resolution, impact of ancillary data, purpose of LULC map and the chosen classifier. Their study also stated the higher classification result of machine learning classifiers than MLC.. Additionally, the study found the importance of including textural information along with spectral data when considering high resolution images for classification. Some other studies compared the effect of training sample size, inclusion of additional spectral bands, pixel and object based classifications on various machine learning algorithms. While classifiers like SVM, Logistic Regression (LR), Logistic Model Tree (LMT) have performed well in lower training sample sizes, RF along with SVM performs well even with complex, high-dimensional data. In the study by Shao & Lunetta (2012), SVM was reported to significantly performed better than Neural Network Classifiers and CART for smaller training samples, SVM also has a superior generalization capability. However a study by Srivastava, Han, Rico-Ramirez, Bray, & Islam (2012) shows a better performance of ANN over SVM in classifying agricultural crops but without a clear reasoning on when such results occur. According to them, more study is required in this direction. With sufficient training samples, most classifiers are reported to perform well (Li et al., 2014). However, with overall accuracy and kappa statistics as the assessment tool in most of the comparative studies, RF and SVM has so far produced higher accuracies than most of the LULC classifiers, even with similar classes. Additionally, they offer low computational cost. These advantages have made RF and SVM as the most widely used LULC classifiers (Jia et al., 2014; Maxwell, Warner, & Fang, 2018; Nery et al., 2016). SVM and RF have used high resolution satellite images to develop higher resolution global land cover maps at accuracy of 64.89% and 59.83% respectively (Gong et al., 2013). Despite these advantages, SVM has the major disadvantage of

being highly sensitive to parameters and defining them is a tedious task. But few studies have shown that another less explored classifier, RVM, overcomes these issues and performs better than SVM for smaller training samples, has reduced sensitivity to hyperparameters, requires less relevant vectors, uses less complex non-mercer kernels and provides probabilistic output. This output can be used to help further increase the classification accuracy (Pal & Foody, 2012).

2.2. Sampling Designs

One of the factors that influence the accuracy of classifiers is the quality of training samples. Obtaining ground truth data for LULC is not feasible most of the times and is an expensive task. Instead different sampling techniques are used to collect training and test data.

Understanding the effect of sampling techniques is important and various existing studies analyse this process.. Stehman (2009) presented, for example, an extensive study on 10 different sampling techniques for accuracy assessment and defined their applicability to different objectives. According to the author, sampling design should be chosen based on objective of accuracy, sampling design criteria and the strengths of the design for the given requirement. Sampling designs discussed in the study include simple random, systematic, stratified random, stratified systematic, cluster random, cluster systematic, stratified random cluster, stratified systematic cluster methods. While most of the studies concentrate on the effect of test data sampling alone on the accuracy of classifications (Stehman, 1992), certain studies shift their focus on understanding the sampling designs for training data selection. For instance, Jin, Stehman, & Mountrakis (2014) investigated different stratified random sampling methods to find how proportional and equal allocation of samples into strata influence the accuracies of classification yielded for urban and non-urban regions. The study further analysed these methods by concentrating the distribution of data within equal sized blocks in each stratum, to understand the effect of spatial allocation. Few studies followed a different approach to define strata for sampling rather than using the class boundaries. The study by Minasny, McBratney, & Walvoort (2007) built a variance Quadtree by decomposing the area of interest into blocks until each of the disintegrated blocks showed equal variability. This is done by taking into consideration secondary variables such as Normalized Difference Vegetation Index (NDVI). This way, observation points are randomly sampled from locations where the surrounding pixels share similar values. Though random sampling and stratification has been the most popular choice for selecting sample points in the remote sensing field, some studies employ systematic sampling methods for land cover studies despite the absence of unbiased estimation of variance.

Systematic sampling generally gives more precise results and hence is generally used in the form systematically generated grids. The most common way for creating grids for larger regions is the confluence of latitude and longitude (Beuchle et al., 2015). Systematic grid sampling which is widely used for soils and in forestry, usually applies statistical techniques like semi-variogram to effectively sample points which provides good estimates of non-sampled locations during interpolation process (Montanari et al., 2012). Such sampling methods were initially discussed in McBratney, Webster, & Burgess (1981) which proposes the use of semi-variogram to create grids with optimal kriging variance. Groenigen & Stein (1998) used Spatial Simulated Annealing (SSA) to optimize the sample distribution where an initial random distribution of sample is moved in random direction and distance 'h' until the distributions reach a state such that mean minimum distance between the sample and non-sample point is reached. This is controlled by the objective function Minimization of Mean Squared Distance (MMSD). This method is proved to be robust and gives

an even spread of samples (Chen et al., 2013). An added advantage of such methods is that they consider the spatial variation of the study area. Such applications prove that geostatistical elements can also help in optimizing the sampling schemes in any area of interest. Gallego (2005) mainly aims to deal with the problem of assigning sample points to one image incase of overlapping images of a large study using Thiessen Polygons and recommends systematic sampling where single points are sampled from these polygons. The study achieves an unbiased estimation of variance in systematic sampling and assigns points of overlapping regions in satellite image frames to the image with the nearest center of Thiessen polygon.

Though there are different studies that focus on sampling strategies for test data, less emphasis is provided on study of these strategies for training data selection (Jin et al., 2014). Understanding the training sampling designs is an important aspect and limited study on this area provides scope for more research (Heydari & Mountrakis, 2018).

2.3. Classification on Google Earth Engine

A continuous attempt to obtain higher accuracy LULC maps always exist. With the availability of free and higher resolution remote sensing images, there are more opportunities for researchers to obtain better than existing maps. This can be achieved using various factors such as choosing the right training samples, including more input features, using multi-temporal higher resolution images, using advanced classification techniques etc. All these contribute to "Big Data" challenge leading to the requirement of extensive computing infrastructure and larger storage space for image classifications (Azzari & Lobell, 2017). According to Giri, Pengra, Long, & Loveland (2013), NASA Earth Exchange (NEX) and GEE help provide a platform to tackle such issues and GEE has expanded as an emerging tool for spatial data analysis.

GEE is a multi-petabyte sized cloud based platform providing parallel computation, data catalogue services for planetary scale geospatial analysis. Computations are automatically parallelized. The public datasets are in ready to use format and ranges from whole United States Geological Survey (USGS) Landsat archives, Landsat Surface Reflectance datasets to Sentinel datasets, various global land cover data, climate datasets and so on. GEE provides various integrated methods which help in pre-processing of images in a simple way. Furthermore, it has a repository of vast functions such as masking, logical operators, sampling data etc., which can be used to perform various operations on images and vectors. Additionally, GEE also allows users to integrate additional logic using Python and JavaScript API. Due to its immense capabilities, GEE has already been used in various LULC based research topics (Gorelick et al., 2017).

Mapping global land cover is an important task in remote sensing. Gong et al. (2013) developed a 30m global land cover map by developing various software on Google Earth that adopts cloud computing. Such analysis can be performed on a single platform, using GEE. For instance, Midekisa et al. (2017) leveraged the power of GEE to produce annual maps for 15 years over the continent of Africa. The global forest cover change map developed by Hansen et al. (2013) uses the power of GEE to obtain multi-temporal 12 year satellite image data and map global forest loss and gain at 30m resolution.

Urbanization is another global issue and there is a need to analyse this change at larger scale. Such large scale LULC classification and corresponding analysis is only possible on a high computational environment. This was achieved through GEE in few studies such as Goldblatt et al. (2016), Trianni, Angiuli, Lisini, & Gamba (2014) and Patel et al. (2015). Such approaches also help in building datasets for national and global scales in a cost effective way. Similarly, GEE has been applied in agricultural sectors such as crop mapping, smallholder farming with a comparative analysis using different in-built machine learning classifiers on a

larger regions with multi-temporal datasets (Aguilar, Zurita-Milla, Izquierdo-Verdiguier, & de By, 2018; Dong et al., 2016; Shelestov, Lavreniuk, Kussul, Novikov, & Skakun, 2017). GEE power has also been used in Digital Soil Mapping (DSM) where it performed 40-100 times faster than the desktop workstation (Padarian, Minasny, & McBratney, 2015).

3. STUDY AREA, DATASETS AND LAND USE LAND COVER CLASSIFICATION SCHEME

3.1. Study Area and Land Use Land Cover Classification Scheme

The study area considered is the region of Dehradun, a district in the state of Uttarakhand, India. Dehradun is located in the northern side of India at 30.3165° N latitude and 78.0322° E longitude, covering around 3088 sq.km of area. It lies in the footlhills of Himalayas and has stretches of Ganga River in the east and Yamuna in the west. Dehradun is the highest producer of fruits in Uttarakhand and contains large spread of plantation and agricultural lands. Dehradun also has a wide coverage of deciduous and evergreen forests which are well protected. The diversity of classes present in Dehradun makes it a good candidate for the study. **Figure 3-1** shows the location of Dehradun District within Indian Boundaries.

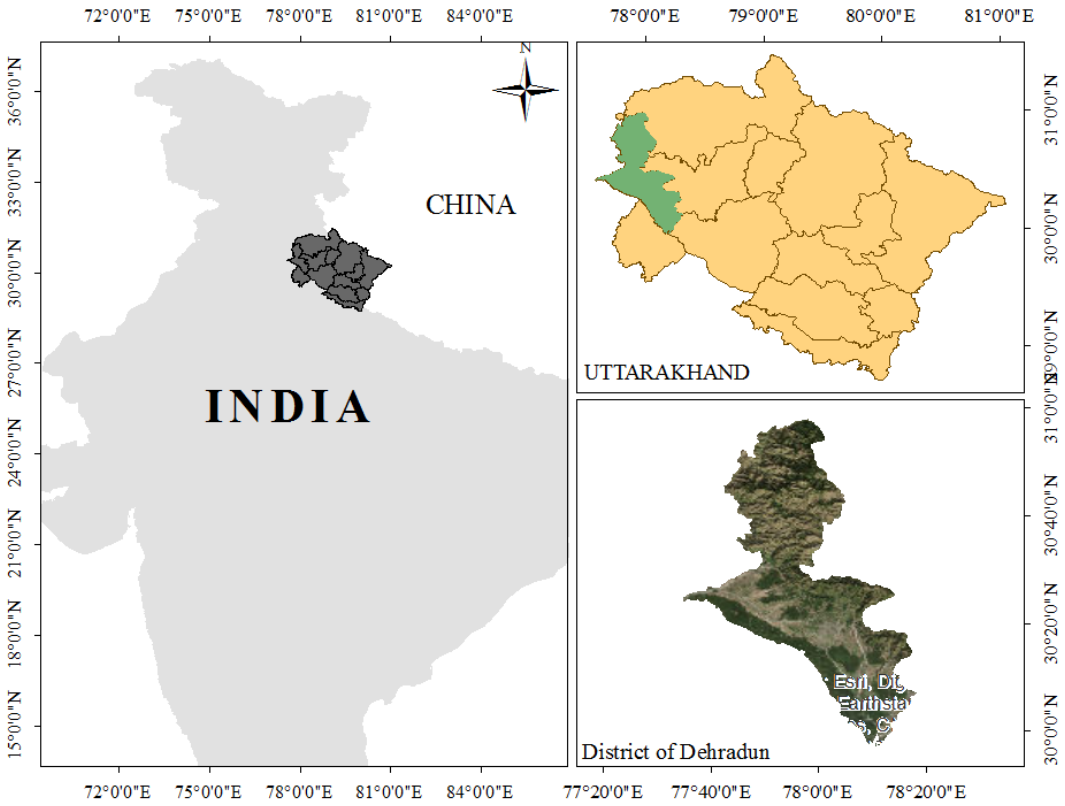


Figure 3-1: Study Area for the research – Dehradun District in Uttarakhand State of India

3.2. Datasets

Datasets for the study are obtained from various sources related to study year of 2017. Datasets can be categorized into two groups based on their purpose in the study – One dataset to perform the classification and another to help identify various classes on the dataset to be classified. Section 3.2.1 and 3.2.2 will describe these datasets further.

3.2.1. Landsat 8 Image Series

USGS provides free repository of medium resolution Landsat image series. The study aims to classify Landsat 8 series data at 30m resolution for the year 2017 to study the LULC of Dehradun . Particularly, the research will use the Landsat-8 Surface Reflectance Tier dataset which are atmospherically corrected using Landsat-8 Surface Reflectance Code (LASRC) and contains 9 bands including two thermal bands.

GEE also has a public repository of freely available data. This includes various satellite images, global land cover maps, water datasets of specific regions, forest cover datasets and so on. The Surface Reflectance Tier 1 data for Landsat 8 directly available in GEE is considered for this study. The year of analysis is 2017 with an aim to concentrate on the latest situation of the study area. The study aims to incorporate the multi-temporal data for the given period. To study classification, various features were selected and two different datasets were built to evaluate the effect of features.

3.2.2. Reference Maps

Unavailability of an LULC Map for the given study area for the year 2017 resulted in using older maps as reference maps to understand the boundaries or strata for various IGBP Classes. Indian Space Research Organization-Geosphere Biosphere Program (ISRO-GBP) 2005 LULC Map, Biodiversity Characterization at Landscape Level (BCLL) 2010 Map, GlobCover 2015 Map were the most suitable maps for the given region and chosen thematic classes. European Space Agency (ESA) released a global land cover map of 300m resolution for the years 1991-2015 under the climate change initiative. The dataset is freely available from the ESA website. ISRO used satellite remote sensing data from 1998-2010 to generate biological richness map for the year 2012. Around 150 land use and vegetation classes were identified and a detailed Biodiversity Characterization at Landscape Level (BCLL) map was generated (Roy, Kushwaha, Murthy, & Roy, 2012).

Table 3-1: Reference Maps for Dehradun LULC Classification

Map	Scale/Resolution
ISRO-GBP LULC 2005	1:250000 Scale
BCLL 2012 Map	1:50000 Scale
GlobCover 2015 LULC Map	300m resolution
High Resolution 2018 Google Earth Images	0.5m resolution

These data sets were selected because reliable land cover classifications have been derived from them using field knowledge, expert consultation and human interpretation. ISRO-GBP reference map contains IGBP thematic classes as required for the study. However, BCLL Map gives data at level 3 category of thematic accuracy and data at level 2 category is required for the study. Hence, different classes were combined into level-2 category to match the IGBP defined classes. The other GlobCover dataset follows Anderson's classification scheme which can be mapped to the classes of interest in the study area. Figure *3-2* shows the

more utilized reference map for the study – BCLL and Globcover over the study area, Dehradun. High resolution images available on Google Earth are used for visual interpretation of the various classes present in Dehradun. These high resolution images are used for identifying the reliability of the above three maps in providing an accurate information of LULC classes for 2017. Section 4.1 and Appendix-A describes how the reference maps are combined to form another reference map for the study.

3.3. Land Use Land Cover Classification Scheme

For the purpose of the study, level 2 thematic classes defined in International Geosphere Biosphere Programme (IGBP) classification scheme is followed. This is also a widely used standard classification scheme in most of the studies (Loveland & Belward, 1997). It contains around 17 land cover classes. Nine IGBP classes are recognized in Dehradun study area and used for classification studies. The classes are: Built-Up, Cropland, Fallow Land, Evergreen Forest, Deciduous Forest, Shrubland, Grassland, Water Bodies, River Bed.

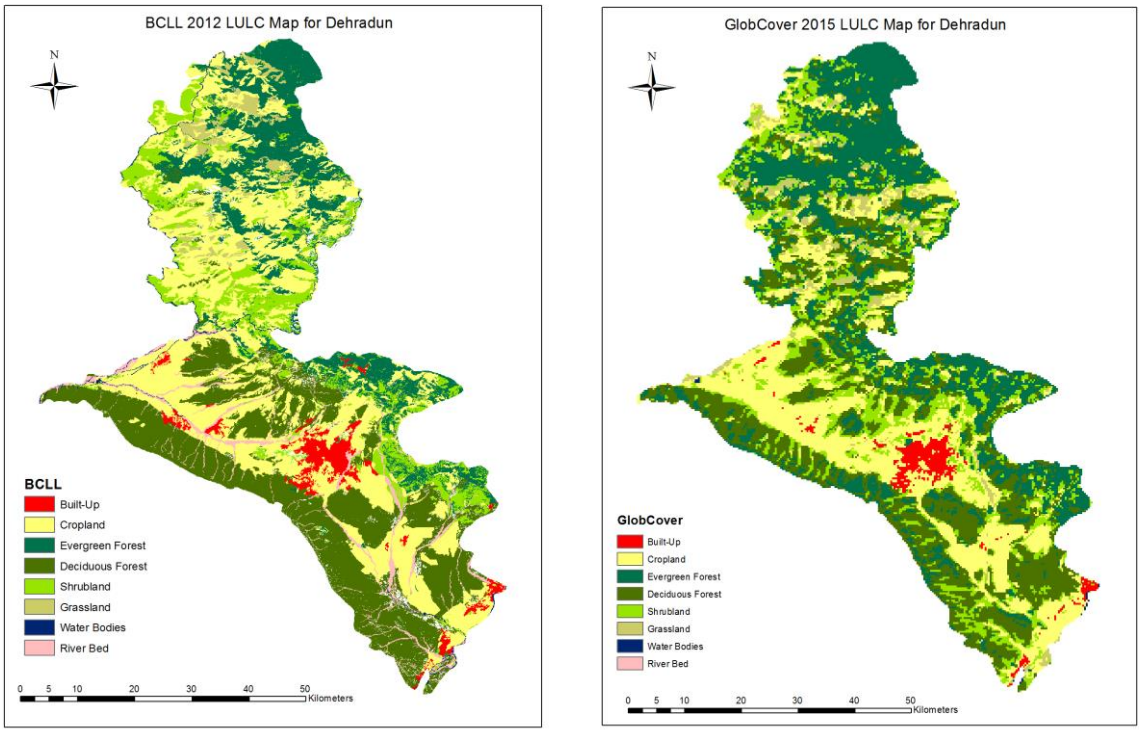


Figure 3-2: Reference maps BCLL 2012 and GlobCover 2015 for Dehradun District

4. METHODOLOGY

This chapter describes the overall flow and various approaches followed to achieve the research objectives. Figure 4-1 depicts the methodology where the whole process is executed in GEE. RVM, which is unavailable in GEE, is implemented in Python outside GEE and GEE Python API calls are made to integrate the two entities together. For certain parts of systematic sampling which involves geo-statistical concepts such as semi-variance and SSA, implementation was performed using R programming language.

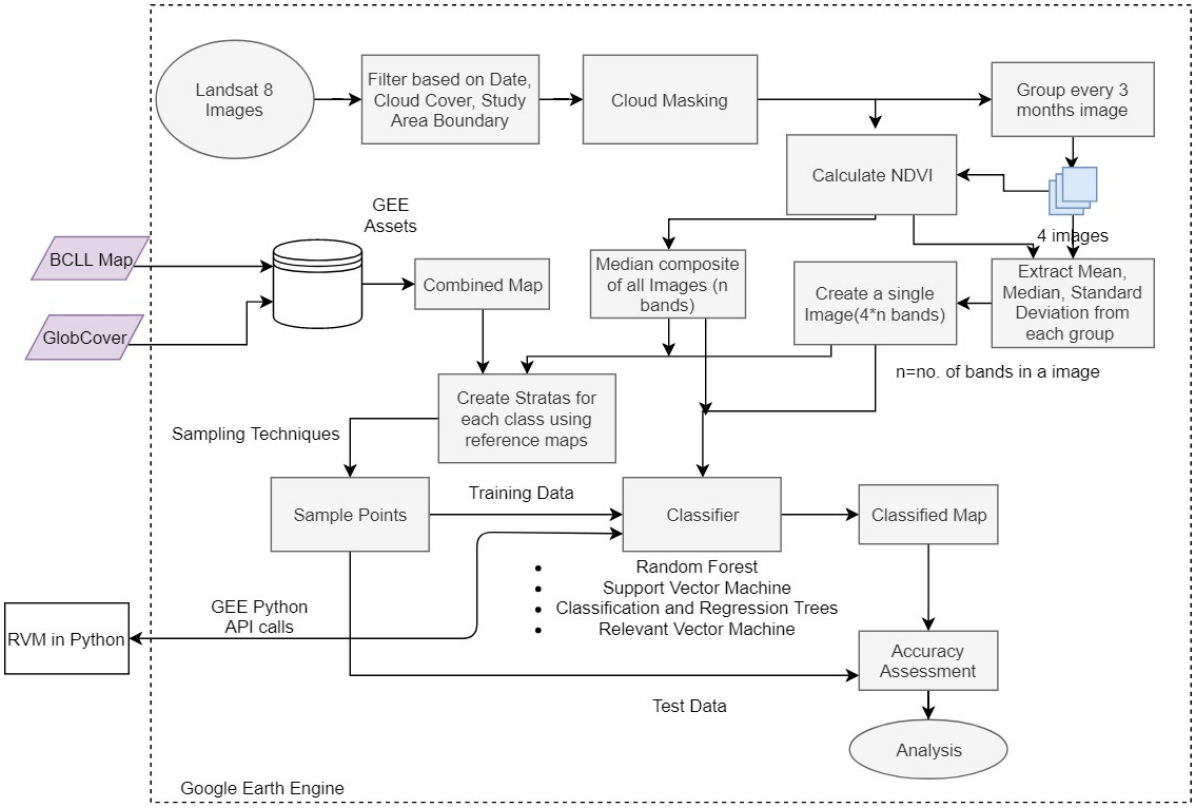


Figure 4-1: Overall Methodology

4.1. Combining Reference Maps

Visual Interpretation of the available reference maps on high resolution Google Earth Images of 2017 indicated changes in land use land cover class present in the reference map and high resolution image. There were also certain dissimilarities of boundaries of LULC classes between the two. These changes will affect the accuracy of the newly classified maps. Additionally, for certain classes GlobCover map provided a better representation of class boundaries on ground and for certain other classes BCLL provided a better approximation. To uniformly assess the accuracy of different reference maps, GlobCover map and BCLL maps are validated using same test data of sample size 100 pixels per class which is defined by visual interpretation using high resolution image, ground knowledge and few field data. Sample size is chosen based on Plourde & Congalton (2003)'s minimum 50 samples per class recommendation. The results of validation are overall accuracy, producer and user accuracy. These class level accuracies help realize the

better of the two reference maps for individual classes. These classes will be extracted from reference maps and combined to create a new map containing class polygons from two different maps. For classes such as Fallow Land whose data are not available for the study area, polygons are delineated manually using ISRO-GBP, for approximately identifying possible fallow lands, and with visual interpretation of high resolution images. Combing all these, gives a new reference map which will be used to define strata for classes from which to sample training and testing data. Accuracy of the combined map will be evaluated using the same test data as described for GlobCover and BCLL.

4.2. Data Preparation

Data preparation for sampling and classification includes filtering data from Landsat-8 collection and integration of multi-temporal changes. The primary set of scaled Landsat 8 Surface Reflectance Scene Tier 1 scene datasets for 2017 is directly obtained from GEE platform. The dataset is further refined to remove cloudy pixels by cloud masking using the available quality bands from Landsat-8. NDVI data is calculated from all the pixels and included as an additional band to the dataset. This is reported to help in further discrimination of land cover classes (Huang et al., 2002). Additionally, since the classes under study mostly contain vegetation classes, contribution of NDVI band is recommended.

To further capture the seasonal and temporal variations, the images are grouped at an interval of 3 months. All images within a 3-month group are aggregated further into a single image containing bands which consists statistical features such as standard deviation, mean and median values for each band. This results in 4 images from each 3-month duration containing 3*N bands (N is the number of bands in an image) in each image. To preserve the statistical data from different periods of the year, all the 3*N bands from each image are combined into a single final image containing 60 bands (4*3*N bands).

To analyse if the above method makes any significant difference in capturing multi-temporal variations, another dataset is created by making a median composite of all Landsat-8 images of 2017. This dataset will also contain lesser number of features (N bands). Table 4-1 summarizes both the dataset.

Table 4-1: Dataset for Image Classification

Data Set 1 (D-1)	Data Set 2 (D-2)
Median Composite of Blue, Green, Red, Near-Infra Red and NDVI bands	Mean, Median, Standard Deviation of Blue, Green, Red, Near-Infra Red and NDVI bands with 3 months image groups

4.3. Stratified Sampling Designs for Selecting Training Data

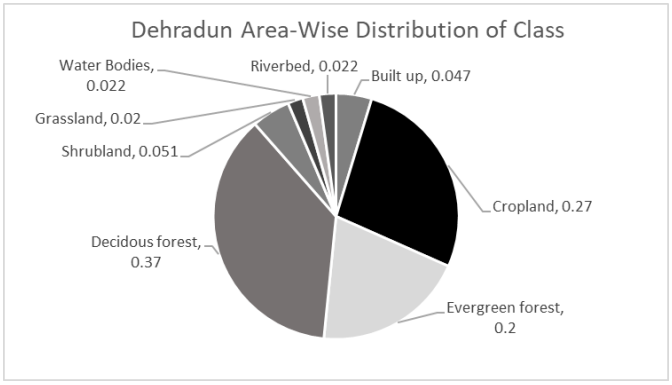
As shown in Figure 4-2 , study area Dehradun contains rare classes such as Grassland, Water Bodies, Shrubland, Riverbed, Built-Up, Fallow Land which cover relatively small portion of the landscape. Other classes such as Cropland, Deciduous Forest and Evergreen Forest occupy relatively larger part of the study area. Stratification ensures that for any given sample size, samples from all classes are considered for training. In the study, Strata are defined for the classes over the image datasets by using the polygons obtained from the reference map. Within each stratum, training data are sampled using random and systematic methods with different sample size allocation in the former, overall resulting in investigation of three different sampling designs – Stratified Equal Random Sampling (SRS(Eq)), Stratified Proportional Random Sampling (SRS(Prop)), Stratified Systematic Sampling (SSS). The sampling units chosen are individual pixels. Effect

of different sampling designs are analysed by evaluating the performance of RF, most well-performed classifier according to various literatures (Belgiu & Drăguţ, 2016), to different training data

4.3.1. Stratified Random Sampling

Sample allocation to strata was investigated under two different methods. While one method allocated the samples proportional to area occupied by each class as per the data shown in Figure 4-2, other method allocated equal samples for all classes. Equal sampling gives unequal inclusion probability for each pixel in every stratum while proportional sampling gives equal probability of inclusion for all pixels in the sample.

Figure 4-2: Area-Wise Proportional Distribution of Classes in Dehradun. Data Source: ISRO Bhuvan



The whole sampling process is performed in GEE on the dataset D-1 and D-2 (Table 4-1). Randomly distributed training samples were obtained using the "stratifiedSample" method available in the earth engine library. This provides flexibility to give different or same sample size for each class, within the area of interest.

To understand the effect of quantity of training samples, different sample sizes were considered to help evaluate a reliable result. Jin et al. (2014) used a small and a reasonably large sample size (1000 and 5000 pixels respectively) in order to understand the effect of training sampling designs. The main aim is to consider both the extremes in terms of time and effort required for sampling data. Following the same approach, this study used Cochran's formula (4.1) for large populations to fix a starting sample size assuming unknown proportion for each class (Cochran, 1953). In equation 4.1, n_0 is the sample size per class, p is the proportion of the population which has the class in question, q=1-p, Z is the z-value for a given confidence, e is the margin of error. For simplicity and maximum variability p is taken as 0.5, e = 0.05 and 95% confidence (Z=1.96). Based on the formula, the sample size was taken as 358 pixels per class. Further, the sample sizes were increased and decreased to capture a recognizable change in classification accuracy. From a total of 24804811 pixels, 0.0001% to 0.0008% of the pixels are overall sampled which are further split into training and testing data.

$$n_0 = {Z^2 pq}/{e^2} \quad (4.1)$$

The effect of two stratified random sampling design for selecting training data is further investigated by evaluating the accuracy of classification using the given training data and its variation with different sample sizes. Test data for validation is randomly selected from each strata and constitutes 30% of the sampled data

4.3.2. Stratified Systematic Sampling

Obtaining training data through systematic sampling has the advantage of choosing samples at a constant distant from each other, thus avoiding similar pixels for training. Fixing on the optimal distance for sampling is an important task. According to Tobler's law of Geography, nearby pixels are more related than distant

pixels. With an intention to choose distinct pixels for training, semi-variance is calculated for each stratum thus optimize the sampling scheme (McBratney et al., 1981).

Considering Z(x) as the continuous variable representing the value at location x, the semi-variogram for a separation 'h' is calculated as half the expectation of difference of values between two locations as shown in equation 4.2

$$\gamma(h) = \frac{1}{2} E\{Z(x) - Z(x+h)\}^2 \quad 4.2$$

Semi-variance gives a picture of the spatial variability of the data. Fitting the calculated semi-variance into a spherical model, helped obtain semi-variogram parameter such as range. Range gives the farthest Euclidean distance between 2 pixels of a class which has maximum autocorrelation and after which spatial auto-correlation ceases to exist. Using range of class/stratum as distance 'd_i' (where i represents different class or strata) for systematic sampling, stratified systematic sampling is performed in GEE on the dataset. For classes that report large ranges, a reduced value is used for initial distribution as large distances do not contribute at all to the final solution (Chen et al., 2013). "sample" method available in earth engine library is used to sample data at specified distances, thus obtaining a set of training samples of different size for each class. To further reduce the distance of separation, SSA is applied on initial distribution of training sample obtained from GEE. Since aim is to find the average minimum distance with which every non-sampled pixel can reach sampled heterogeneous pixels, Minimization of Mean Squared Distance (MMSD) objective function is used. SSA and MMSD are implemented using spsann package of R developed by Samuel-Rosa, Heuvelink, Vasques, & Anjos (2015). The initial temperature is set to high values in the range of 7000 – 50000 so that the perturbations in the large areas be accepted at 95% rate in the first chain (Chen et al., 2013) and the process continues till a stopping condition is reached while cooling the temperature. The MMSD obtained for all classes at the end of the annealing process will be used as distance of separation for systematic sampling. Final sampling of data is performed on GEE using the newly obtained distance to form a systematic spread of samples. The advantage of this method is that the sample size is automatically determined by the spread of selected training data.

SSS sampling aims to select heterogeneous pixels from a class such that they represent most of the variations of a class. Rather than the relying on the chance of randomness for choosing the right training samples, SSS chooses samples based on the variations of data within a class and by placing training samples such that they are at a distance which represents minimum variation within a class. It aims to select limited heterogeneous training samples which can represent the variation of a class.

4.4. Classification using in-built classifiers

GEE platform provides complete support to perform various image classification processes. The machine learning classifiers of the current study i.e. RF, SVM, CART are in-built in GEE and are used for training the classifier with the samples obtained from various sampling designs.

4.4.1. Classification and Regression Tree

CART is a binary decision tree classifier which takes simple decisions for logical if-then questions. The classifier examines the input variables and the chooses the variable with maximum information gain based on which the node splits at every level (Breiman et al., 1984). In this technique, the input data is randomly divided into certain number of groups and trees are generated using all the groups leaving out one. The left

out group is used for validating the tree and the pruned tree which gives the least deviance is taken. In this study, classification using CART was performed in GEE using the Classifier.cart method available in Earth Engine library. The best choice of cross validation factor is 5 or 10 according to studies (Kohavi, 1995) and is considered as input value.

Table 4-2: Input Parameter values for CART Classification

Cross Validation Factor for Pruning	5 & 10
Maximum Depth of the Initial Tree	10
Minimum training points to allow creation of a node	1
Minimum training points at a node to allow splitting	1

4.4.2. Random Forest

Random forest is an ensemble of k CART classifier which overcomes the overfitting problem of CART. It is the most widely used classifier for LULC Classification and hence a classifier of interest for the current study. Random forest applies bagging technique and randomly selects subset of features from input observations for each tree. The main input parameters for RF are the number of trees and the variables at each split. Very large number of trees does not necessarily increase the accuracy of classification, because after a specific number of trees, additional trees do not contribute more in the prediction of labels and become redundant. Studies suggest 100 or 500 as the optimal number of tree count and square of number of variables as an optimal count of variables to decide the best split (Belgiu & Drăguţ, 2016).

Random Forest classification is performed on GEE using the Classifier.randomforest method available in earth engine library. Table 4-3 shows the input parameter values taken for analysis for the current study.

Table 4-3: Input parameter values for Random Forest

Number of Trees	50,100,150,200
Number of Variables per split	Square root of input variables

4.4.3. Support Vector Machine

SVM is another widely used classifier which works based on finding an optimal hyperplane that separates the decision boundary between different classes. The selection of the support vectors mainly depends on the choice of cost parameter C, Gamma and kernel functions. The cost parameter decides the level of punishment for a misclassified data. Higher the value of C, lesser the misclassified data within a class

The parameters are chosen through a grid search method to obtain a good C and Gamma which gives accurate prediction results. According to studies such as Hsu, Chang, & Lin (2003), exponentially growing sequence of C gives better approximation to good parameter selection as C is a scale parameter. Additionally, for large datasets linear kernel is preferred for training (Hsu et al., 2003). The gamma parameter is not valid for linear kernels. Following these approaches, the choice of parameter for the current study is shown in Table 4-4. The Classifier.svm from earth engine library is used to implement SVM in GEE.

Table 4-4: Input Parameter Values for SVM

Kernel Type	Linear
Cost Parameter	2^{10}, 2^{11}, 3510, 2^{12},2^{13},2^{14},2^{15}
SVM Type	C_SVC

4.5. Relevant Vector Machine and Google Earth Engine Integration

RVM is another Bayesian classifier which is reported to give similar accuracy results as SVM with the added advantage that it overcomes all the shortcomings of latter. RVM with a new fast marginal likelihood marginalisation developed by Tipping & Faul (2003) is a faster version of original patented RVM by Tipping (2001). This potential RVM classifier, which is not available in GEE is analysed for LULC classification by implementing and integrating the overall classification task between GEE and local hardware as shown in Figure 4-3 .

.

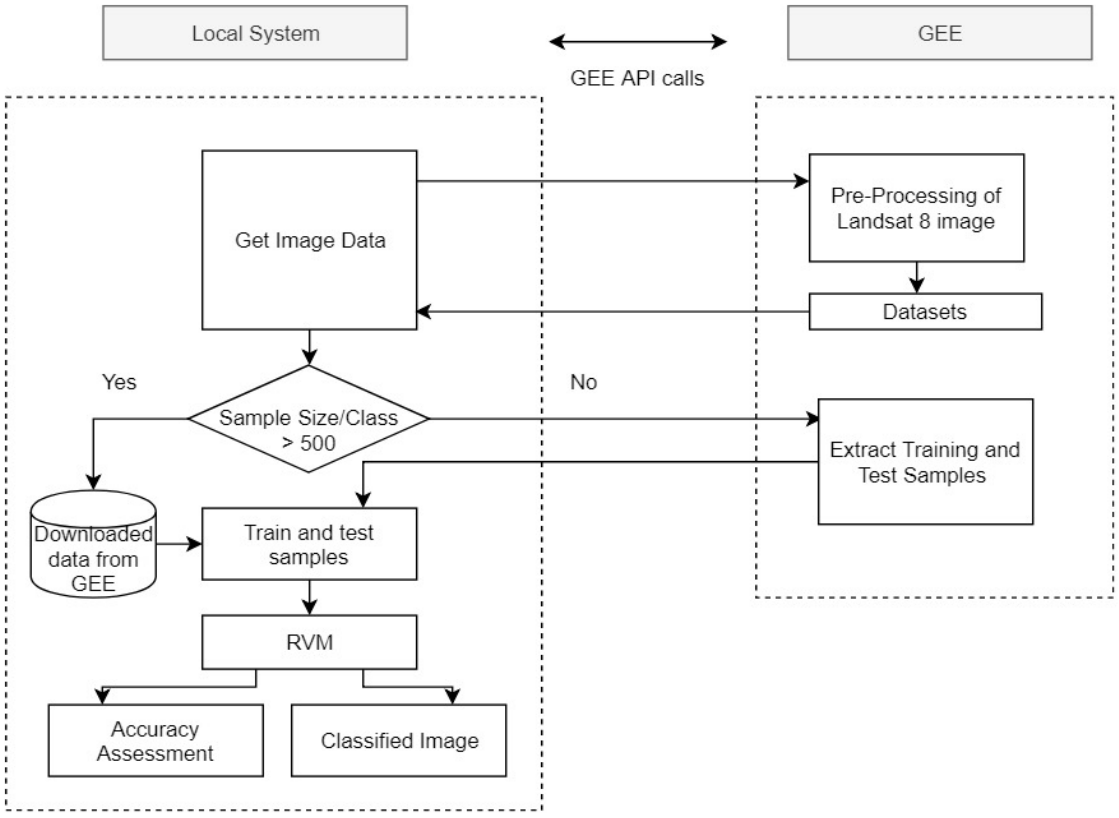


Figure 4-3 :Roles of Local System and GEE during RVM implementation. The intensive processing tasks of creating multi-temporal datasets are performed on GEE.

While main RVM classifier logic is implemented in python, pre-processing of satellite images to create classification datasets and employment of sampling techniques to extract training and testing data requires

GEE. GEE allows interaction of an external python program with Earth Engine servers. This is achieved with the help of following set of packages provided by GEE,

- google-api-python-client – helps to authenticate to earth engine servers
- earth-engine-api - earth engine python library to make API calls to execute python program in GEE and to import GEE results or datasets.

RVM classifier is trained in the local system based on the training set and performs accuracy assessment using test samples obtained from GEE.

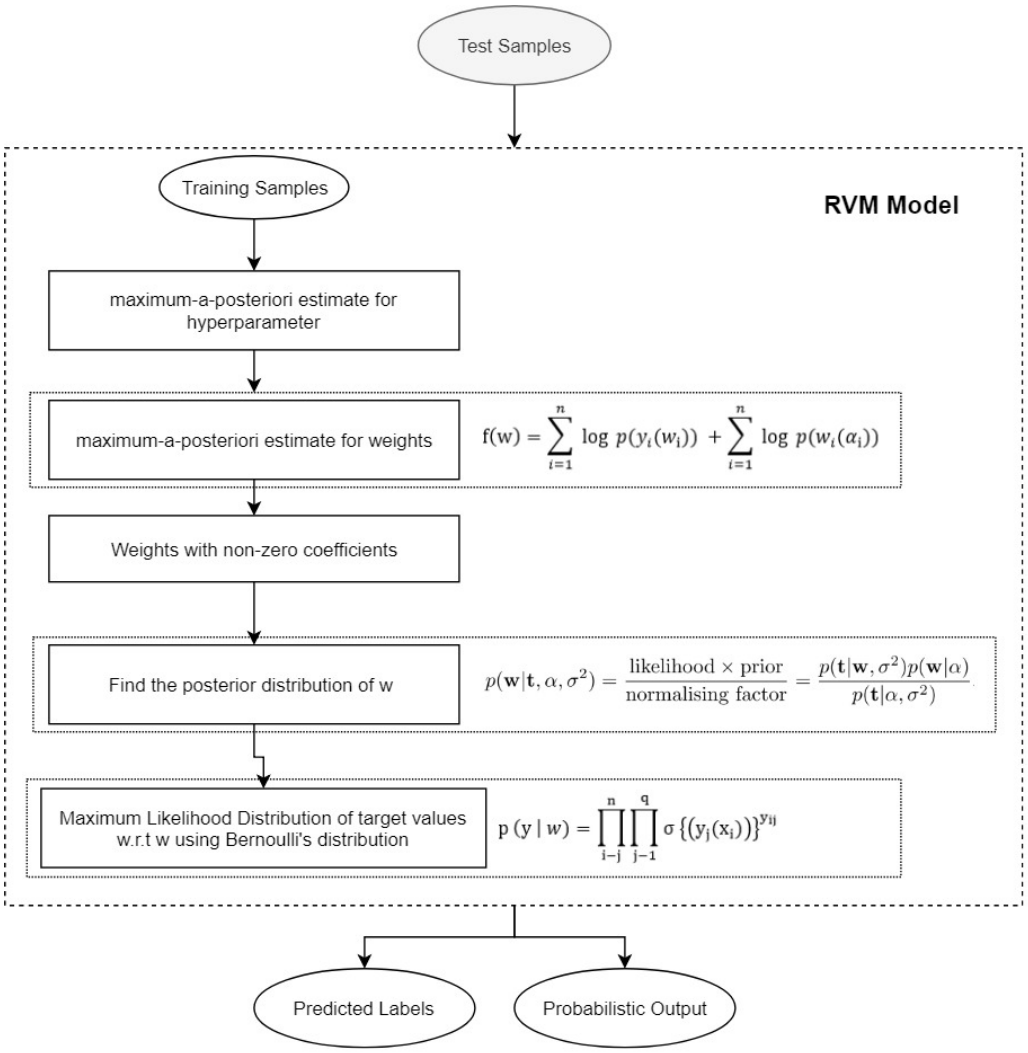


Figure 4-4: Logical Flow for Relevant Vector Machine

A verified python implementation of RVM is available in github (Shaumyan, 2017) which provides a *sklearn-bayes* package for RVM. Further RVM is implemented in python mainly using *Numpy, Scipy and scikit-learn* libraries. Figure 4-4 explains the logic of the implemented RVM algorithm. Based on the input training data, the Bayesian classifier tries to build a model with an overall aim of finding the probability distribution of

target values (y) based on the variables (w) present in the training data(x). Since the model tries to minimise the error by fitting a probability distribution which finds the best relation between the given target value t and weight distribution w, the model can become too specific to the given data leading to generalization errors. This is controlled by introducing a prior in terms of hyperparameter α. Solving equation 2.5 gives a variable set w which has mostly larges value of hyper-parameter associated with it. Such values will tend to zero and will not be considered further in building the model. This induces sparsity in algorithm making use of only smaller set of relevant vectors for training. With the availability of variance σ^2 for given samples, prior probability of w, target value t and hyper parameter α, posterior probability distribution of w is found. The parameters for t are estimated with maximum probability using Bernouilli likelihood as shown in equation 2.3. The model is executed with different dataset and sample size with input parameter being the choice of kernel required with which to map training data to a higher dimensional feature space.

4.6. Accuracy Assessment of the Classification Results

Accuracy assessment helps evaluate the performance of various classifiers and also the effect of the underlying training sampling designs. In the study, test samples are randomly chosen from each strata defined over the dataset using reference map such that they don't overlap with the existing training data which are created using various sampling designs. For stratified random sampling designs studies same test data has been used for a given sampled size of training data to assess different classifiers. To summarize, if A represents the total number of training and testing data, 30% of A represents the count of test data.

Among the various available metrics, Overall Accuracy (OA) is used for assessing all classifiers performance as well as the effect of sampling designs through the performance of one of the classifiers i.e. Random Forest. Overall Accuracy is the most widely used metric which helps in easy interpretation and is effective in accuracy estimation (Plourde & Congalton, 2003). It expresses the percentage of test data which has been correctly classified by the classifier. Additionally, with an intention to further evaluate the class-level performance of a given classifier, confusion matrix, user accuracy and producer accuracy are used (Stehman, 2009). GEE provides methods to achieve the accuracy assessment for different classifiers such as errorMatrix, producersAccuracy, consumersAccuracy and accuracy.

Another aspect of focus is to understand the performance of RF, SVM, CART and RVM classifiers when compared to each other. The relative comparison is performed using Z-Score as described in literatures such as Congalton & Green (2010), Foody (2009), Rossiter (2014). The comparison shows if there is any significant difference between the performance of the classifiers. Let p_1 and p_2 denote the proportions of the correctly classified test data of the two classifiers of interest while s_1 and s_2 represent the standard deviation of their samples. With the assumption of independent sample distribution, the significance of comparative results of two classifiers is given by equation 4.3.

$$Z = \frac{|p_1 - p_2|}{\sqrt{{s_1}^2 + {s_2}^2}} \qquad 4.3$$

Given the null hypothesis, H_o: $|p_1 - p_2| = 0$ and alternative hypothesis H_1: $|p_1 - p_2| \neq 0$, Z value is calculated for a given confidence level $\alpha/2$ of two-tailed Z test and null hypothesis is rejected if Z is greater than or equal to $Z_{\alpha/2}$. With 95% confidence, equation 4.3 is used to compare the resulting maps of two different classifiers. If the corresponding z-score is greater than 1.96, then the comparative results suggest with more than 95% probability that one classifier is better than the other. For all the classifiers, results are obtained from same sampled training and test data for comparison, thus avoiding any bias in analysis.

5. RESULTS AND ANALYSIS

5.1. Accuracy of Reference Maps

Accuracy of any LULC map is highly influenced by underlying datasets, reference map and level of thematic classes to some extent. Reference Maps are an alternative to ground truth data which can be used to generate labels for training and test samples for classification. Since the current study had a large study area of 3088km^2, collecting ground truth was out of scope. In the absence of an accessible reference map for 2017, reference map which is closest to the year of analysis and thematic accuracy should be chosen. Appendix-A shows the results of user and producer accuracy of individual maps. The results of the accuracy individual GlobCover, BCLL maps and combined maps are shown in Table 5-1 . The results were tested on test samples obtained from field data and visual interpretation of high resolution images. The results show an improved overall accuracy of 83.66% for the new combined map as against 71.2% of BCLL map and 67.50% of GlobCover map.

Table 5-1: Overall Accuracy of reference maps validated on test sample of size 100/class

BCLL Map (2012)	GlobCover Map (2015)	Combined Map
71.2%	67.50%	83.66%

5.2. Effect of Sampling Design on Training Data

This sub-section presents the results for various sampling designs implemented to obtain training data. Training data is a subset which represents the study area and the classes present in it. To assure that both small and large classes are considered in training data, three different stratified sampling methods are analysed – SRS(Eq), SRS(Prop), SSS. Impact of sampling designs are analysed further by evaluating their performance in RF classifier.

5.2.1. Stratified Random Sampling

Table 5-2 shows the overall accuracies of RF classifier that is trained using training datasets obtained from two different stratified random sampling techniques – SRS(Eq) and SRS(Prop). SRS(Prop) method produced better overall accuracy when compared to SRS(Eq) method. The result follows a similar pattern when applied for different sample sizes (2250, 3222, 9000, 18000 pixels) and datasets (D-1, D-2). For a given sufficiently large sample size 9000 which includes pixels from all classes in good amount for training, SRS(Prop) performs at an average 7.12% better than SRS(Eq). For smaller sample size of 3222, there is no significant difference between overall accuracy for dataset D-1 and D-2. The effect of training samples in relation to the size of a class can be evaluated through producer and user accuracies as shown in Figure 5-1

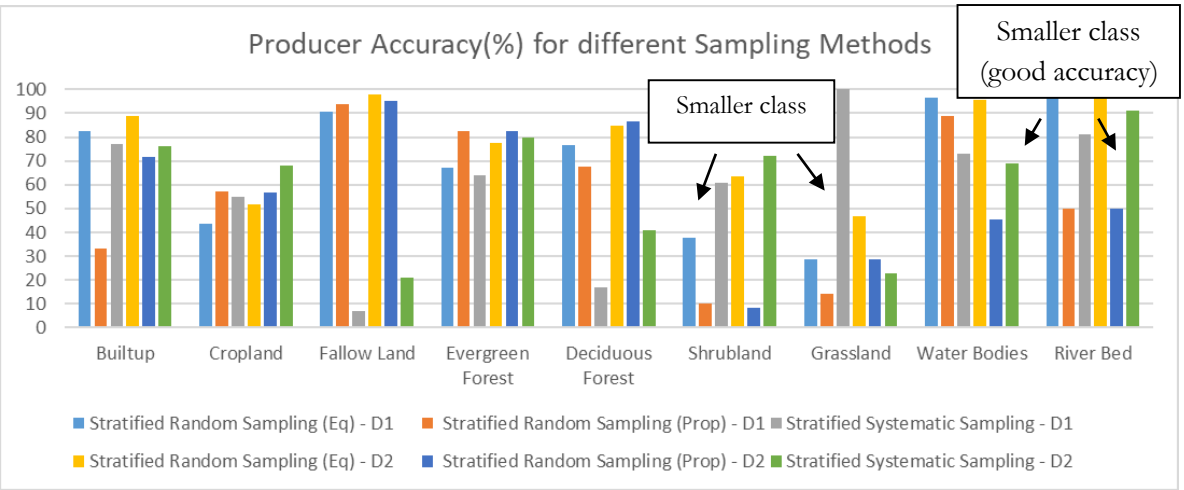


Figure 5-1: Producer Accuracy for different sampling methods obtained by RF Classification on a sample size 3222 for Stratified Random Sampling methods and a sample size of 8460 for Stratified Systematic Sampling method

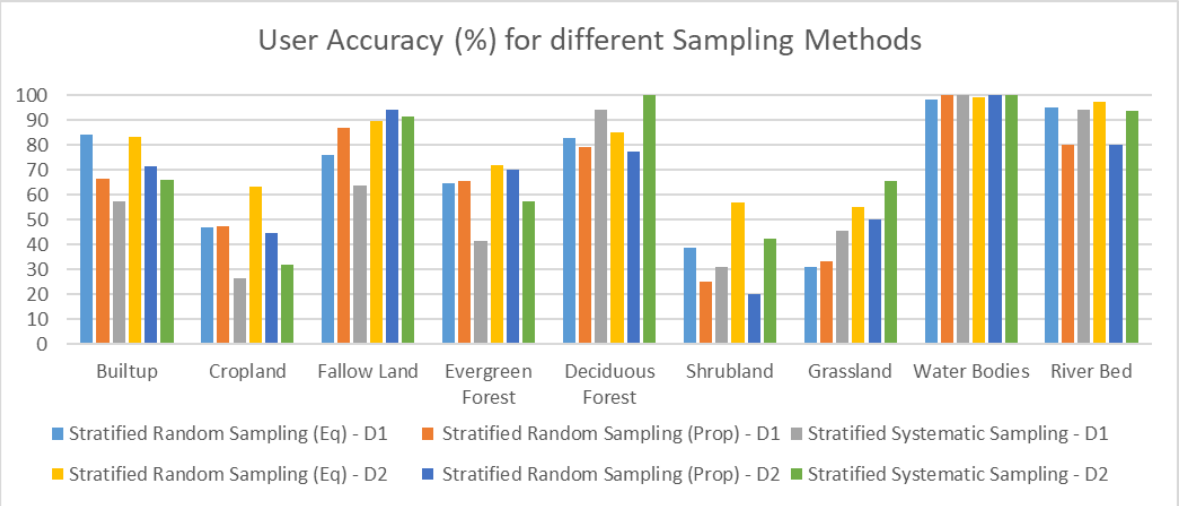


Figure 5-2: User Accuracy for different sampling methods obtained by RF Classification on a sample size 3222 for Stratified Random Sampling methods and a sample size of 3800 for Stratified Systematic Sampling method

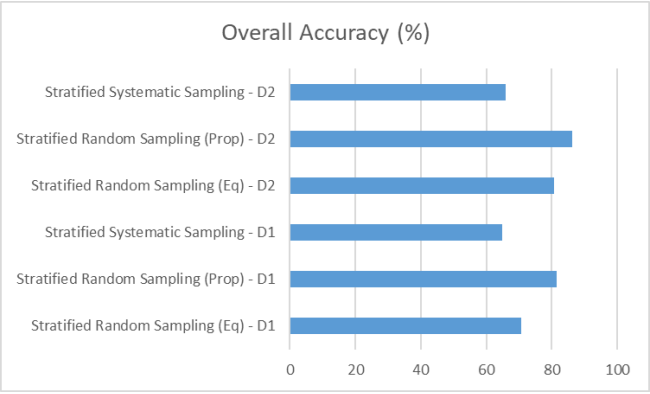


Figure 5-3: Overall Accuracy of different sampling methods validated using similar sample size

Table 5-2: Accuracy of Random Forest Classification for Stratified Random Sampling Methods. The table shows results for a given sample size and two different datasets D-1 and D-2. The training and testing sample set size is same for all classes in SRS(Eq) method. For SRS(Prop), sample size depends on the area occupied by the class in the study area.

	Stratified Equal Random				Stratified Proportional Random		
RF Classifier Parameter	No.of Pixels/Class		Accuracy (%)		Total Pixels all class	Accuracy(%)	
No. of Trees	Training	Testing	D-1	D-2	Training+Testing	D-1	D-2
100	250	108	71.25	78.93	3222	75.2	76.42
100	700	300	71.21	79.38	9000	80.48	86.14
100	1400	600	70.56	80.61	18000	81.58	86.35

and Figure 5-2. Larger classes such as Evergreen Forest, Deciduous Forest and Cropland which occupy 37%, 20% and 27% of the training samples in case of SRS(Prop), have much better producer and user accuracy than those of smaller classes such as Shrubland, Grassland which contain 5.1%, 2% of the training samples. In contrast, smaller classes, including Built-Up which has 4.6% of the training sample, have significantly higher user and producer accuracies in case of SRS(Eq) method with an average increase in performance of 2.07%-55.33% from SRS(Prop). The improvement can be attributed to difference in sample distribution count between SRS(Prop) and SRS(Eq). 11.11% of training samples are present in each class in case of SRS(Eq) which is much more than the amount of training samples obtained for smaller classes by sampling in SRS(Prop) method. On the other hand, larger classes with more samples in SRS(Prop) perform slightly better (1.5%-15%) than their corresponding SRS(Eq) results. This shows that smaller classes have more advantage on using SRS(Eq) method than larger classes.

Certain smaller classes such as Water Bodies, Fallow Land and River Bed which represent 2%, .89%,2% of the training samples in SRS(Prop) have user accuracies ranging from 76.11% to 100% and producer accuracies of above 90% for both SRS(Eq) and SRS(Prop) methods. This can be explained by spectral distribution graph which is discussed in Section 6.1.

5.2.2. Stratified Systematic Sampling

SSS requires decision on the sampling distance between the pixels based on certain criteria. SSS was started with an initial sample design where each class pixels were separated by a distance of "Range". Range is calculated based on semi-variance of each class and the variogram parameters obtained are shown in Appendix-B. The initial sampling of data with Range values gave an overall low accuracy of 36.65% (More details in Appending-B). The results of applying SSA on initial distribution sample points of each class to systematically obtain a new minimum distance using objective function MMSD is presented in Table 5-3. The number of samples in the initial distribution is automatically determined due to range criteria for separating the samples. Range values vary from 216m for Fallow Land to 3637m to Evergreen Forest. Classes with less intra-class variability such as large Evergreen Forest and Deciduous Forest classes have larger Range values. Interestingly, small class of Shrubland has a large range of 2220m. This can be attributed to small variability within Shrubland class. Initial Sample distribution with such large range is irrelevant as it contains very less training samples (of around 5) for a large study area. The result obtained may not be reliable for further processing has such samples will not be a good representative of the class. Additionally, SSA+MMSD for such large range will take lot of time which can be avoided. Hence a reduced value of 1500

is taken for initial distribution of these three classes. The automatic determination of sample size can be an advantage in this process.

Table 5-3: Range values obtained from semi-variogram model and Minimum Mean Squared Distance obtained by SSA using MMSD objective function on the initial sample distribution of each class. For classes with very large Range values such as Evergreen Forest, Deciduous Forest and Shrubland, initial sample separation is not the same as Range. For such classes, the initial sample distance taken is additionally mentioned. The MMSD value is used for systematic sampling which will automatically determine the sample size.

Class	**Range (m) / Initial Separation**	**Minimum Mean Squared Distance Separation (m_**	**Final Training Sample Size**
Built-Up	320	280	1006
Cropland	504	363	2719
Fallow land	216	185.69	450
Evergreen Forest	3637/1500	259.44	1092
Deciduous Forest	2440/1500	1056.965	377
Shrubland	2220/1500	242	2415
Grassland	333	250	207
Water Bodies	1200	314	63
River Bed	1110	279.23	131

Table 5-4: Error Matrix for RF classifier on Stratified Systematic Sampling with 100 validation pixels per class. The results are for dataset D-2. [BU – Built-Up, CL – Cropland, FL – Fallow Land, EV – Evergreen Forest, DE – Deciduous Forest, SL – Shrubland, GL – Grassland, WB – Water Bodies, RB – River Bed]

		REFERENCE MAP								
CLASSIFIED MAP		**BU**	**CL**	**FL**	**EV**	**DE**	**SL**	**GL**	**WB**	**RB**
	BU	**73**	6	6	0	0	1	4	9	5
	CL	16	**72**	72	6	9	20	53	5	9
	FL	0	3	**3**	1	2	1	0	1	2
	EV	0	4	4	**79**	11	8	3	0	0
	DE	0	1	1	3	**62**	0	2	0	0
	SL	1	12	12	11	15	**68**	14	6	0
	GL	0	2	2	0	1	2	**24**	0	0
	WB	0	0	0	0	0	0	0	**72**	2
	RB	10	0	0	0	0	0	0	7	**82**
	Overall Accuracy: 66%									

Application of SSA+MMSD has drastically reduced the initial separation value for some classes such as Water Bodies (reduction of 73.8%) and River Bed (reduction of 74.84%). Built-Up, Cropland, Fallow Land and Grassland have seen a reduction of 12.5%-27.97% from the initial sampling range distance. Classes with large range have a significantly reduced value of separation distance between samples from MMSD. With

new value of separation between the sample obtained, systematic sampling is performed for each class. The overall accuracy of training RF classifier with the SSS training sample is 64.89% for D-1 and 66% for D-2. In Table 5-4, quantity of correctly classified pixels and the confusion between various classes can be observed for SSS method on D-2. The matrix shows good classification of large forest classes and a high confusion between Cropland and Fallow Land. Certain amount of overlap also exists between spectrally similar River Bed and Built-Up classes. Shrubland has been misclassified into 7 of the rest 8 classes. The confusion between water bodies with cropland and shrubland indicates poor quality of reference map used for identifying the actual labels of training samples, as water bodies is spectrally distinct from other classes.

When compared to SRS(Eq) and SRS(Prop), the producer and user accuracy of SSS for all classes is consistently good. Though in most cases, SRS methods have produced better results, the accuracies obtained from SSS are still reliable. From Figure 5-3 we can observe that at similar sample sizes, SRS(Prop) performs 16.69% better than SSS while SRS(Eq) performs 5.67% better. The pattern is similar for different datasets considered.

5.3. Relevant Vector Machine in LULC Classification

The results of applying RVM for LULC classification on 2 datasets D-1 and D-2 are shown in Table 5-5. The table shows that RVM gives highest overall accuracy of 64.01% for the smallest sample size of 175 samples/class considered in the study. For the same sample size and another dataset D-1 with lesser number of features, the accuracy of RVM decreased by 10.29%. As the number of training sample increases, the accuracy of RVM decreases. Given a dataset D-2, the accuracy decreased by 5.43% for an increase in sample size of 125 training pixels per class. As the sample size increased to larger 700 training samples/class (6300 sample size), the accuracy decreased at a much slower rate than those for smaller training sample sizes.

Table 5-5: Overall Accuracy Results for Relevant Vector Machine Classification on datasets D-1 and D-2. Each row represents results for different sample sizes. Training and Test samples are split such that they form 70% and 30% of the sample respectively in each class.

No.of Pixels/Class		Accuracy(%)	
Training Samples	Validation Samples	D - 1	D - 2
175	75	53.72	64.01
250	108	50.44	58.58
700	300	49.94	55.67

Among the given training samples, RVM selects only smaller subset for training the model and classifying the data. These subset samples are called relevant vectors which have non-zero coefficient associated with its weights. Table 5-6 shows the number of relevant vectors for each class of 175 training pixels. The useful training sample size forms only 0.011%-0.017% of the initial sample size. This constitutes very small training sample set and observes a huge reduction in the sample size required for classifying the data.

Table 5-6: Count of chosen relevant vectors per class for different initial training sample size. The results tabulated are for dataset D-2

Sample Size/class ↓	Builtup	Crop Land	Fallow Land	Evergreen	Deciduous	Shrubland	Grassland	Water Bodies	River Bed
175	3	2	2	2	2	2	2	3	2
250	2	2	2	2	2	2	2	2	2
700	2	2	2	3	1	2	2	2	2

Table 5-7: Error Matrix for Relevant Vector Machine Classification on D-2 data set for sample size of 175 pixels/class. The number of test samples are at an average of 75 pixels/class. The class level performance in terms of producer accuracy (PA) and user accuracy (UA) for all 9 classes as calculated from error matrix is depicted under PA row and UA column. [BU – Built-Up, CL – Cropland, FL – Fallow Land, EV – Evergreen Forest, DE – Deciduous Forest, SL – Shrubland, GL – Grassland, WB – Water Bodies, RB – River Bed]

		REFERENCE MAP									
CLASSIFIED MAP		**BU**	**CL**	**FL**	**EV**	**DE**	**SL**	**GL**	**WB**	**RB**	**UA (%)**
	BU	**36**	6	0	1	3	0	3	8	19	*47.38*
	CL	4	**24**	9	6	11	10	5	1	4	*32.43*
	FL	2	1	**66**	0	4	0	0	0	1	*82.19*
	EV	0	2	2	**47**	15	5	2	2	0	*62.67*
	DE	0	5	0	8	**52**	8	13	0	0	*60.47*
	SL	1	3	12	21	13	**24**	3	1	1	*30.38*
	GL	0	7	14	17	19	13	**4**	1	0	*5.33*
	WB	1	0	0	3	3	0	3	**142**	8	*90.45*
	RB	4	2	2	0	0	0	0	3	**160**	*93.57*
	PA(%)	*75*	*48*	*62.85*	*45.63*	*43.33*	*40*	*13.33*	*90.45*	*82.47*	**555**

Table 5-7 which gives a count of correctly classified and misclassified pixels by RVM classifier in the error matrix. Out of the 867 pixels, 555 pixels are classified correctly using the chosen 20 relevant vectors. Classes such as Water Bodies and River Bed perform well with more than 90% user accuracy. Water Bodies is also confused with River Bed to certain extent. Most of the Cropland classified pixels belong to either deciduous or evergreen forest. The fallow land pixels are misclassified as Shrubland and Grassland classes to certain extent that it reduced the producer accuracy to 62.5%. Large Deciduous and Evergreen Forest classes have low producer accuracy of 45.63% and 43.44% respectively due to their confusion with Shrubland and

Grasslands. User accuracy of Built-Up is at 47.38% mainly due to its overlap with River Bed. In most of the vegetation classes, the accuracies are reduced due to Shrubland and Grassland. This can also indicate the low performance of RVM in identifying distinct pattern in green vegetation classes.

RVM's output gives an additional information about the posterior probability distribution of a test sample into various classes during classification. This information will help in identifying the most erroneous class or the class that creates most of the confusion. Table 5-8 presents the posterior probability distribution of misclassified test samples based on the probability values of the classified classes (highest class probability assigned to a sample). The probability values are distributed into quartiles and each quartile shows the count of pixels incorrectly allocated into a class. 1st quartile ranges from 0.173195 to 0.29783 and contains maximum number of misclassified pixels i.e. 133 of the total 312 incorrectly classified pixels. Among the 133, deciduous forest alone gets incorrectly labelled into 50 times. Other classes such Built-Up, Cropland, Evergreen, Shrubland and Grassland are also incorrectly allocated labels. Second quartile which has 124 misclassified pixels also has a lower posterior probability of final incorrectly allocated class. Among these, Evergreen forest and Shrubland form the major allocation with 24% and 19.35% misclassified respectively. Towards the higher quartile 3, the posterior probability value increases but the number of misclassified pixels' decrease. Misclassifications with probability more than 42% is observed only in a subset of classes such as Fallow Land and Evergreen Forest to a larger extent and Shrubland, water bodies and river bed to a comparatively smaller extent. Quartile 4 shows the highest probability misallocated pixels counts which amounts to 14 pixels. Fallow land and River bed are the only classes which are misclassified with high affinity of 7 incorrect pixels respectively towards them. The highest reported probability value is 0.671733 and indicates the maximum confidence with which the classifier misclassifies a pixel. The last row of Table 5-8 shows total count of incorrectly allocated pixels into each class and indicates forest classes to be misclassified into maximum number of times. Built-Up and Water Bodies show the lowest misclassifications with 12 and 15 pixels respectively.

Table 5-8: Misclassified test pixel count distribution based on posterior probability quartile of final classified/allocated classes. [BU – Built-Up, CL – Cropland, FL – Fallow Land, EV – Evergreen Forest, DE – Deciduous Forest, SL – Shrubland, GL – Grassland, WB – Water Bodies, RB – River Bed]

Quartile	**Posterior Probability**		**Misclassified Pixel count in allocated classes**									**Total Mis-classification**
	Min	**Max**	**BU**	**CL**	**FL**	**EV**	**DE**	**SL**	**GL**	**WB**	**RB**	
Q1	0.173195	0.29783	12	18	8	13	50	10	12	5	5	**133**
Q2	0.29783	0.422465	0	8	10	30	18	23	14	6	15	**124**
Q3	0.422465	0.5471	0	0	14	13	0	3	0	4	7	**41**
Q4	0.5471	0.671733	0	0	7	0	0	0	0	0	7	**14**
Total misclassification/allocated class			**12**	**26**	**39**	**56**	**68**	**36**	**26**	**15**	**34**	*312*

5.4. Classification Results of Machine Learning Classifiers

The in-built machine classifiers of GEE were applied for LULC classification of the study area based on the training samples obtained by stratified equal random sampling method. Another classifier of interest, RVM, which was implemented outside GEE but integrated with it for complex processing was also used for LULC classification of the same study area. The results of these classifications are shown in Table 5-9 for the best possible input parameter selection. The results are evaluated on different datasets D-1 and D-2 and sample sizes. Performance of the evaluated classifiers show an increase in overall accuracy when dataset D-2

containing more features is used. For instance, CART shows an average increase of 6.54%, RVM shows an increase of 8.31%, SVM shows the lowest increase of 3.99% and RF shows the maximum average increase of 8.94%.

With the variation in sample size, the classifiers show a variation in performance trend. While SVM and RVM gave a classification accuracy of 69.63% and 64.03% respectively for the smallest sample size of 175 pixels/class, their accuracies decreased by 6.09% and 8.87% respectively for the larger sample size of 1400 pixels/class. Similar result can be observed in dataset D-1 with similar performance trend of SVM and RVM towards increased sample size. Whereas, CART and RF classifiers performed similarly with increase in sample size with overall accuracies concentrated around 72% and 79% respectively. A clear understanding of the same can be obtained from Figure 5-6 which shows the graph of RF and CART classifier as a near

Table 5-9: Overall Accuracy of CART, RF, SVM, RVM for different sample size and datasets D-1, D-2.

No. of Pixels/Class		**Overall Accuracy(%)**							
Training Samples	**Validation Samples**	**D - 1**				**D - 2**			
		CART	**RF**	**SVM**	**RVM**	**CART**	**RF**	**SVM**	**RVM**
175	75	66.28	71.51	64.13	53.72	72.08	79.58	69.63	64.03
250	108	64.46	71.25	64.46	50.44	72	78.93	66.06	58.58
700	300	66.44	71.21	58.09	50.22	72.57	79.38	64.12	55.67
1400	600	65.62	70.56	57.98	50.91	72.99	80.61	60.84	55.16

straight line graph. while RVM and SVM show a steeper decrease within smaller training samples size of 1575 and 2250 pixels, than between the small sample size of 2250 and larger sample size of 6300. The graph also shows the consistent higher performance of RF compared to other three classifiers.

RF performs better than other classifiers even at class level accuracies as shown in results of producer and user accuracy in Figure 5-4 and Figure 5-5 respectively, with an exception of very few cases. For instance, Water bodies showed a slightly higher accuracy of 98.17% for SVM compared to 95.63% of RF. SVM

Table 5-10: Z-statistical test for 95% confidence showing Two-Tailed Probability of 1st classifier performing better than 2nd classifier. Classifier results were obtained for a test sample size of 675 pixels on dataset D-2.

Classifiers	**Z-Score**	**Confidence**
RF vs SVM	4.2272	99.9976
RF vs CART	4.3403	99.998
RF vs RVM	6.4459	99.998
CART vs RVM	2.1889	99.8546
SVM vs RVM	3.1836	97.1396
SVM vs CART	0.99081	67.8216

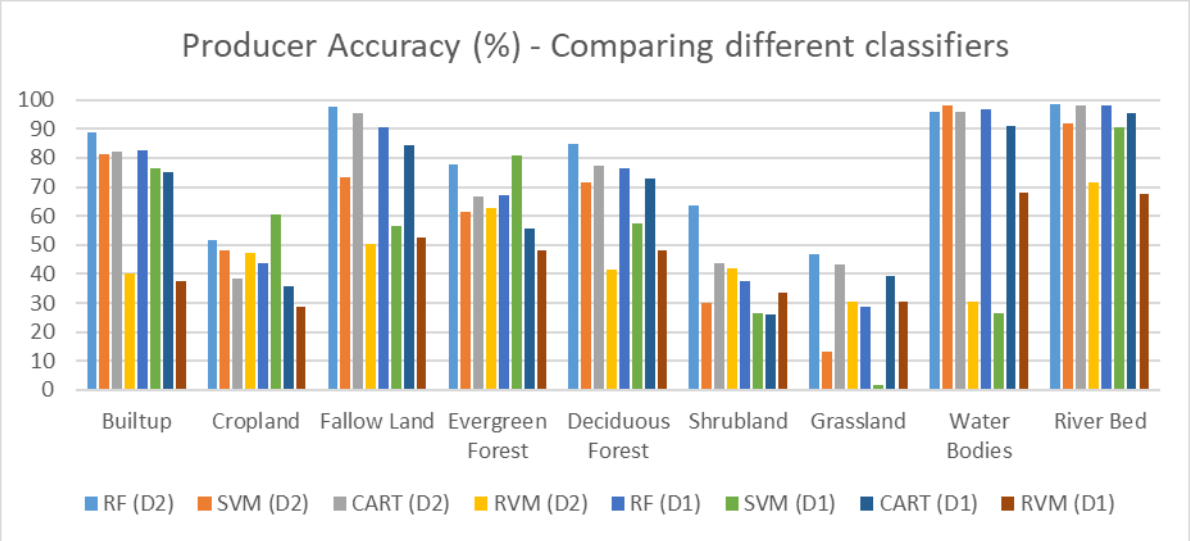


Figure 5-4: Producer Accuracy for different classifiers with a sample size of 6300 pixels

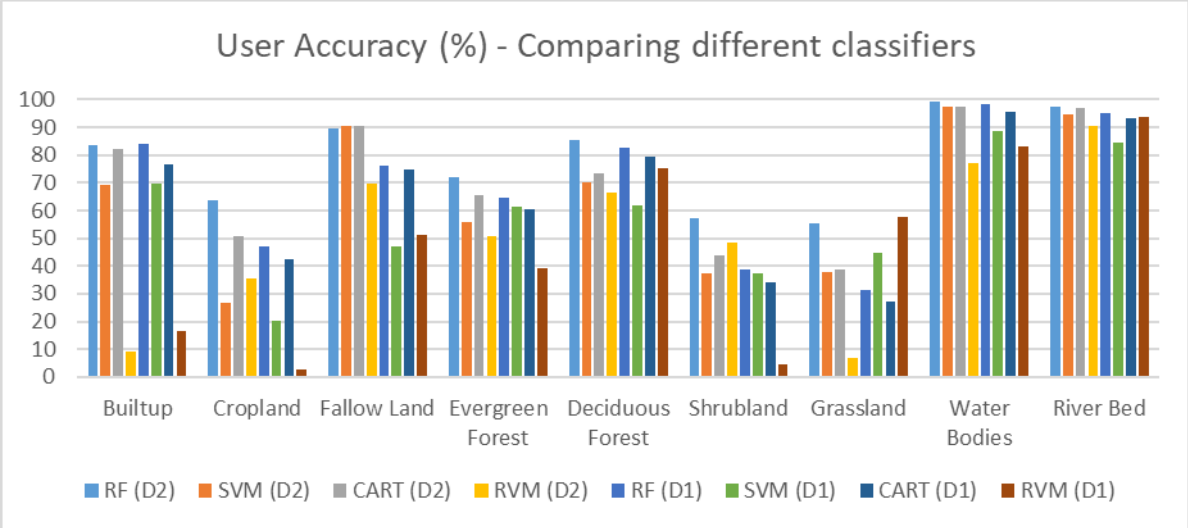


Figure 5-5: User Accuracy for different classifiers with a sample size of 6300 pixels

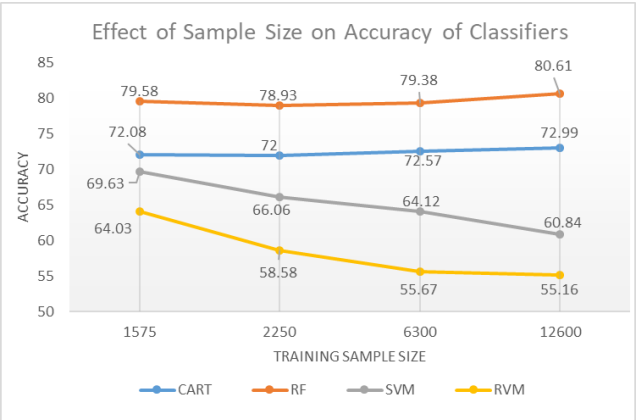


Figure 5-6: Change in accuracies of classifiers to change in training sample size

performed better with higher feature set D2 than D1 except in Evergreen Forest class. Shrubland and Grassland are low performing classes in all four classifiers, where SVM performed the lowest in 62.5% of the cases (Among the 8 various combinations of producer and user accuracy with dataset D1 and D2, SVM

performed lowest in 5 of the cases). CART performs second best user and producer accuracy results next to RF 72.22% of the time. Rest 27.78% of the scenarios are when SVM is performing better RVM on the hand, gives relatively low producer and user accuracies in most of the cases. Class-level performance of classifiers can be further analysed from the spectral distribution of different bands involved in the classification as discussed in Section 6.1 Figure 5-7 shows the LULC map of Dehradun district obtained from CART, RF, SVM and RVM which helps visualize the distribution of classified classes in different methods.

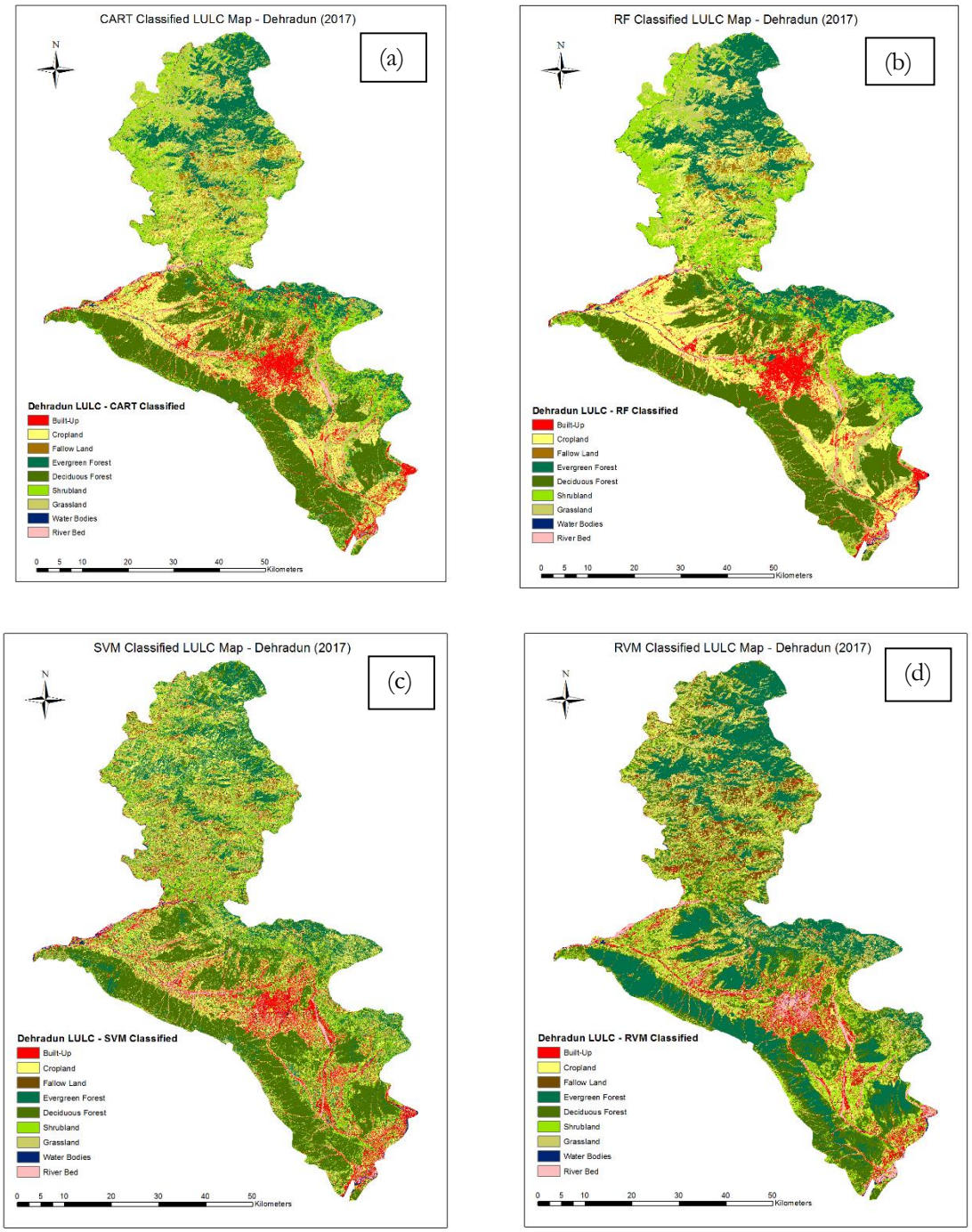


Figure 5-7: Classified Map of Dehradun District using different machine learning classifiers (a) CART (b) RF (c) SVM (d) RVM

To affirm the significance of difference in performance between various classifiers considered in the study, Z-test is performed which is depicted Table 5-10. At 95% confidence, all the z-scores above 1.96 indicate the significantly higher performance of one classifier over the other. With the exception of SVM vs CART, all the classifier combinations showed statistically significant result. RF performs better than all classifiers 99% of the time, SVM performs better than RVM with 97% probability, CART performs better than RVM with more probability than with SVM. The results of Z-test between CART and SVM is not significant because the z-score is 0.99 which is less than z-score of 1.96 for 95% confidence.

6. DISCUSSIONS

Obtaining an accurately classified LULC map is of great importance in remote sensing field. Such accuracies are affected by various factors such as choice of classifiers, quality of training data, heterogeneity of the landscape, dataset, reference maps and so on. This study analyses the different sampling methods and important machine learning LULC classifiers such as CART, RF, SVM, RVM to understand their impact on LULC classification using multi-temporal Landsat 8 image series.

6.1. Impact of Reference Maps and Datasets

In the absence of an accessible LULC maps for 2017, reference maps which are closest to the year of analysis and thematic accuracy are chosen. The study considered ISRO-GBP (2005), GlobCover (2015) map, BCLL (2012) map and visual interpretation using very high resolution Google Earth Images to define reference polygons for various classes. The accuracies of individual maps were less and certain classes were better defined by particular maps when compared to the other. Selecting references of different classes based on producer and user accuracies helped build a more accurate LULC reference map for the year 2017. This resulted in new reference map performing on an average 15% better than individual maps. Integrating few field visit data along with test data defined by visual interpretation provided more reliability to the accuracy of the reference map. Since more accurate reference maps are always desired, this approach can be followed to obtain reference maps where the required reference data is unavailable. However, the reference map contains errors in terms of mislabelled location due to which the classifiers in the study may not produce their best possible results. These errors can also be inferred from the spectral profile distribution of training samples as shown in Figure 6-1 obtained by labelling points based on reference map. There are overlaps of certain classes such as Shrubland and Grassland with most of the other classes, which has brought down the accuracies of classifiers discussed in the study. This is also in-line with the studies by Foody et al. (2016) who found through intentional introduction of misclassified training pixels into reference maps, the reduction in accuracy and performance of classifiers. Still, since same reference map and samples are used throughout the current study, the results can be reliably interpreted and analysed. The study also takes advantage of presence of incorrect reference data to understand the sensitivity of classifiers to training sample quality.

The dataset used for classification should always be analysed to understand how it can be best utilized to obtain an accurate LULC representation of the remotely sensed images. The difference in accuracy can be observed in various results obtained in Chapter 5 for dataset D-1 and D-2. From multi-temporal Landsat-8 image series of 2017, dataset D-1 extracts only the median composite of surface reflectance values for blue, green, red and Near Infrared (NIR) bands. This does not capture the data about variation of dynamic classes such as cropland, shrubland, grassland, deciduous forest, water bodies and river bed within a year. For instance, the cropland of Dehradun District consists of plantations, pulses, cereals, food grains and each of these crops show a variation in growth, colour, density during the different seasons (Tuteja, 2013). Few agricultural lands might also remain bare for some time after the harvest season, but still needs to be labelled as cropland. Other classes such as Water Bodies change their boundaries during summers and increases in volume during rainy seasons. This inversely influences the formation of River Beds. Another example of variation is that of Deciduous Forest shedding their leaves on the onset of summer. These reasons make dataset D-2 a more favourable option as it captures the multi-temporal changes provided by the data along with variations within a group (mean, median, standard deviation of bands values). The change in behaviour of various classes in different seasons can be observed in Figure 6-1 and Figure 6-2

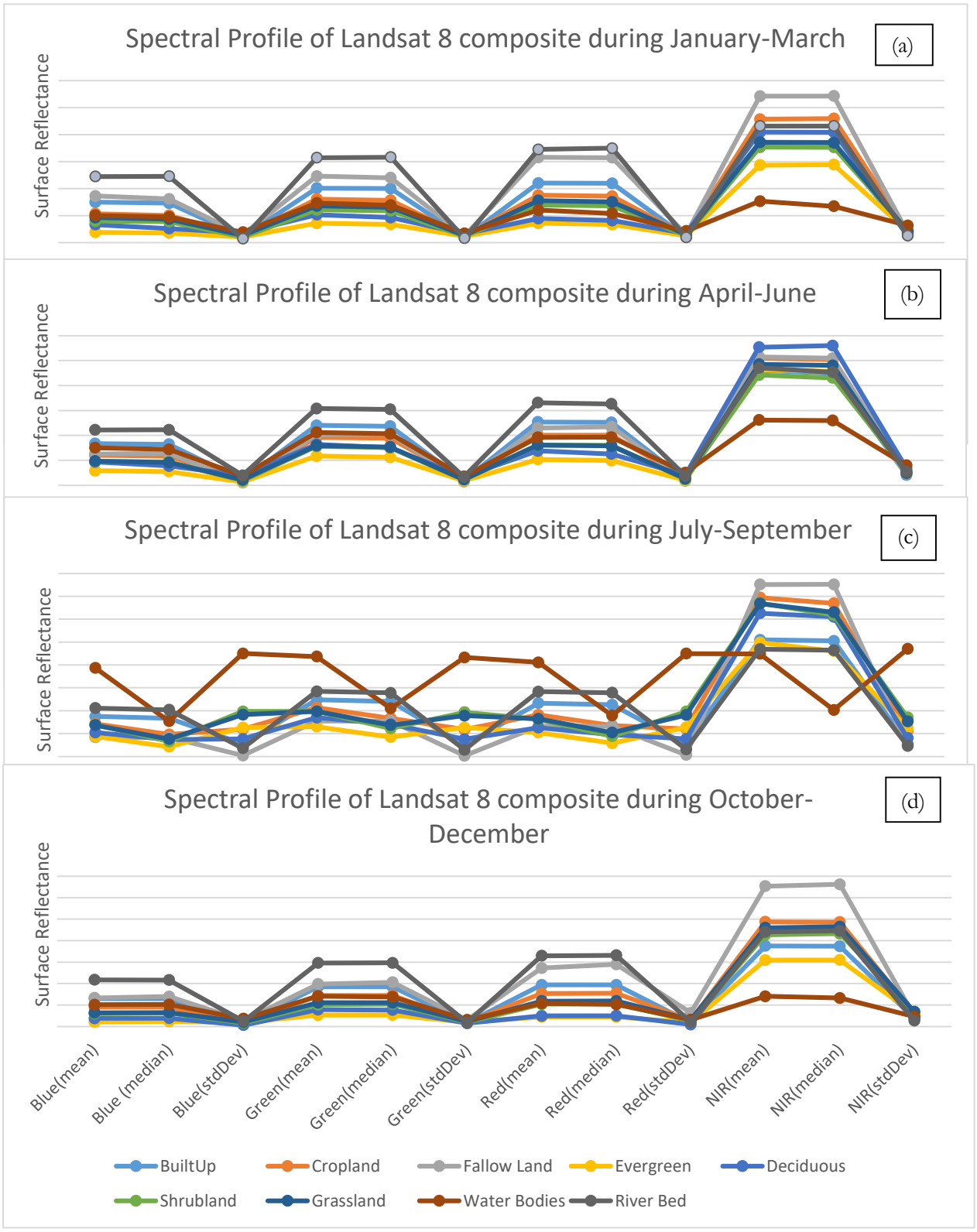


Figure 6-1 (a-d): Variation of sample values in dataset -D2 for different months of 2017

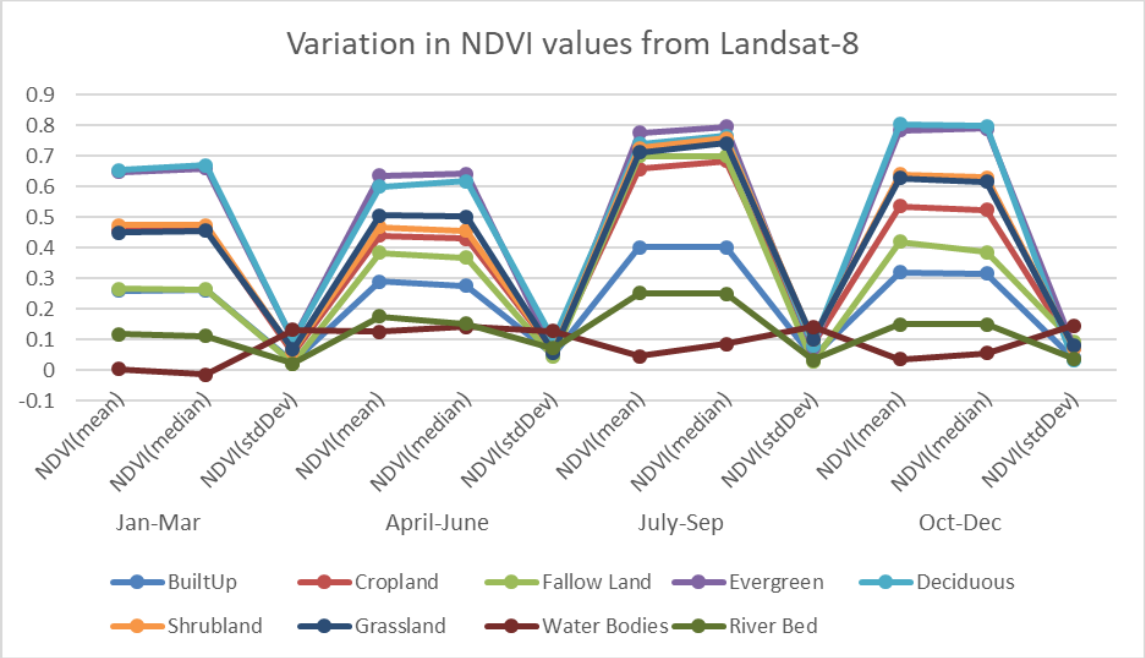


Figure 6-2: Variation of NDVI data in D-2 for different months

Figure 6-1(a) shows River Bed and Fallow Land well separated from rest of the classes in all the bands while Cropland and Water Bodies overlapping near the blue band. April-June months see strong overlap of cropland and fallow land due to dry crop lands reflecting similarly to fallow lands. During the same period, shrubland shows overlaps with different classes making it less distinct (Figure 6-1(b)). Certain overlaps could also be attributed to the incorrectly labelled training samples as discussed. While all groups show a certain trend of statistical values, classes in the months of July-September show largest overlap and dissimilar pattern with an exception in NIR region and for Water Bodies (Figure 6-1(c)). These months show larger values of standard deviation indicating larger variability of classes within this period. Such periods could also be further divided to capture more accurate variations. Removal of this data might help in achieving better accuracy results. The last 3 months show similar class discrimination as the first 3-month composite but the NIR band does not discriminate the classes well in (Figure 6-1(d)). Among all the considered bands, Figure 6-2 shows that NDVI band discriminates classes the best in all months due to the presence of mostly forest and vegetation classes in the study. Throughout the series, Shrubland and Grassland showed largest overlap with other classes.

Hence an understanding of the study area can help prepare relevant datasets for classification. But for larger study areas where understanding the nature of all minute classes can be a difficult task, a generic dataset such as D-2 that approximately groups images based on defined global seasonal months and captures statistical variations, can be used for classification.

6.2. Impact of Sampling Methods on Training Samples

Stratified sampling helps include data from all classes. This assures that even smaller classes are considered in sampling. The sampling method to be chosen depends on which accuracy feature is important in the study. If the aim is to get an overall good accuracy of LULC maps, SRS(Prop) method can be used for

sampling. SRS(Prop) is an equal probability sampling where all pixels of a class have equal chance of being chosen in the random sampling process. However, smaller classes such as Grassland, Shrubland are under-represented in SRS(Prop) resulting in poor performance at class level (Figure 5-1 and Figure 5-2). On the other hand, SRS(Eq) allocates equal number of samples for small and large classes (with unequal probability for pixels within a class to be chosen). This results in smaller classes such as Grassland and Shrubland being represented well. SRS(Eq) sampling gives equal weightage to all classes in terms of sample count and might over-represent smaller classes resulting in less generalization. But overall class-level performance for SRS(Eq) is better than SRS(Prop) as even the smaller classes perform well due to good representation in terms of size in samples. This outcome is also according to the study published by Jin et al. (2014) who compared stratified random sampling with equal and proportional sample allocation method for urban and non-urban classes. Another study by Colditz (2006) on evaluating different training sampling scheme for tree-based classifiers found that allocation of training samples according to area of each class produces best results

While Stratified Random Sampling methods depend on the randomness in which samples are picked and the allocation schemes (Equal or Proportional), SSS is a non-random process which considers the underlying variation of sample variables measured using semi-variance for determining the sampling points. This method can under-represent a class if only Range of fitted semi-variogram model was considered to determine the sampling distance. But the application of SSA using MMSD helps identify the optimal distance for sampling, thus optimizing the sampling scheme. Unlike SRS(Eq) and SRS(Prop), SSS has lesser tendency of over-representing or under-representing a class. This is because the size of training samples in each class is controlled based on the MMSD obtained. Additionally, MMSD helps in placing training samples such that they are at a distance which represents minimum variation within a class. This can be evidenced from Figure 5-1 and Figure 5-2 where SSS has performed well even for Grassland which showed poor performance in other methods. The comparative performance of different sampling methods shows the consistent performance of SSS sampling in all classes irrespective of the proportion of each class in the study area. While SRS(Eq) and SRS(Prop) favour certain classes based on their sample size, SSS fits well into all types of classes irrespective of their class area size. SSS is suited for classes with great amount of intra-variability such as such as Grassland or large classes such as deciduous and evergreen forest where proportional and equal sampling can over-represent a class. While sample size selection is at the freedom of the user as per the requirement in case of SRS(Eq) and SRS(Prop) methods, SSS limits the training sample size based on the underlying variation which can be an advantage as well as a disadvantage based on the requirement. Unlike in Van Groenigen & Stein (1998) where SSA+MMSD is applied to place the randomly distributed sample points in a systematic grid, this process is used in the current study to find the optimal separation between sampling points. But this can also create a risk of error propagation if the first randomly chosen training sample is erroneous. Table 6-1 summarizes the various advantages and disadvantages of sampling techniques studied for training sample.

Figure 5-1 and Figure 5-2 shows smaller classes such as River Bed, Water Bodies and Fallow Land which perform really well in all methods. This can be attributed to two reasons – The distinct values of chosen training samples (could be due to less variability within a class random sampling assures training sample well represents the class) and the accuracy of reference map for each class. The reference map defines the three classes with good agreement when validated using few field data and visual interpretation using high resolution images. further shows the distinct values of River Bed and Water Bodies when compared to other classes. Additionally, larger Cropland class which occupies 27% of the training sample is expected to perform better for SRS(Prop) than Built-Up. But overlap of cropland and fallow land during the summer months of April-June contributes to confusion of dry cropland with fallow land, thus reducing its performance (Figure 6-1(b)).

Table 6-1:Summarized advantage and disadvantages of different sampling methods for training samples

Stratified Equal Random Sampling	
Advantages	Disadvantages
• Flexibility to choose any sample size • Simple and Efficient sampling process • Avoids under-representation of smaller classes	• When the number of classes are large, large sample size is required to get good representation of each class. This could be computationally expensive
Stratified Proportional Random Sampling	
Advantages	Disadvantages
• Good overall accuracy compared to other methods • Flexibility to choose any sample size	• Smaller classes are under-represented
Stratified Systematic Sampling	
Advantages	Disadvantages
• Automatic Sample size determination • Considers variation of dataset values within a class	• Not much control on number of samples • Less applicable for homogeneous classes • Chance of Error Propagation with erroneous first pixel

6.3. Analysis of Relevant Vector Machine

RVM classifier for LULC performed better with smaller training sample sizes. An accuracy of 64.01% was achieved with a sample size of 175 pixels per class. The performance even decreased with the increase in sample size. This shows the inducing of additional complexity in the learning process which makes it more confusing for the classifier in finding the right relevant vectors among the many for training the model. At larger sample sizes, the increase in accuracy is not significant. This shows the relevance of RVM for smaller training samples. Further, for a given sample size of 175 pixels per class RVM finally trained the classifier with a very small proportion of training sample, i.e. using only 2-3 vectors per class. This accounts for significant reduction in training sample count. The results are similar to the study by Pal & Foody (2012) where classification of 6 crop types involved only 1-10 relevant vectors. Similar reduction in useful training samples are seen for higher sample sizes but at the cost of decrease in accuracy. This quality of RVM which chooses only those training samples with non-zero co-efficient obtained from finding the posterior probability of hyper-parameters, makes RVM an attractive feature in remote sensing field where training samples/ground truth are limited.

RVM classifier shows most of the misclassifications in vegetation classes such as Evergreen Forest, Deciduous Forest, Cropland with Grassland and Shrubland. This could be attributed to the fact that only two relevance vectors represented the large classes which could not distinguish the spectrally similar classes of forest and vegetation. RVM performs comparatively better for smaller classes. Hence RVM is more applicable for classification on smaller study areas.

The probabilistic output of RVM can be studied to understand the problematic classes. It gives a picture of uncertainty in allocation of classes for classified test samples. This property of RVM is another important feature which can be used in identification and analysis of sources of errors for all cases of misclassified pixels. Finding sources of error can help improve the classification accuracy as the user is aware of the problem inducing classes. For example, among the misclassified pixels, most of the classes are classified with highest probability to Fallow Land and River Bed. Further analysis indicates that most of the time Built-Up class is misclassified into River Bed. Fallow Land has been a source of confusion to many classes such as Shrubland and Grassland. Based on such analysis, further refining of input data can be performed by adding more features which will help distinguish such confusing classes better. RVM chooses relevance vectors away from the boundary, unlike SVM. Such training pixels should be well separated in feature space from other class pixels. But since the results indicate a confusion spectrally different class such as Fallow Land and Shrubland/Grassland, it confirms the presence of incorrectly labelled training samples within these classes. This anti-boundary nature of RVM makes it more sensitive to training data quality.

The posterior probability distribution of misclassified pixels to allocated class shows low value of probability for majority of the pixels. On the other hand, only 14 out of the 312 are misclassified with high probability indicating the classes which misclassify with confidence and needs further refining to achieve better accuracy. Similar results have been obtained by Foody (2008) where there are 7 confidently misclassified pixels out of the total 320 misclassifications. The study later identifies these 7 pixels and associated classes to be major reason for classification error. Thus the probabilistic nature of RVM can be utilized to study the misclassified classes and potentially use the information to refine the classification output, training-testing samples and so on.

6.4. Comparison of Machine Learning Classifier Performance

The results of machine learning classifiers are influenced by many factors. One of the factors is the features used for training from the underlying dataset. Same classifier produces different results for training and test sample points taken from same locations when the underlying dataset of Landsat-8 is different. The difference in accuracy is observed in Table 5-9 where RF, SVM, CART and RVM performed better with dataset D-2 which contains more number of features than in dataset D-1. Additional features are reported to improve the classification accuracies for machine learning classifiers as long as sample size is more than number of features Maxwell et al. (2018). Though our study saw an increase the performance of all classifiers, RF and RVM could take maximum advantage of these additional features as their performance improved by 8% from D-1 to D-2. Studies by Alonso, Malpica, & Agirre (2011), Du, Samat, Waske, Liu, & Li (2015) show such results for RF, SVM and CART but the effect on RVM has not been studied so far.

Though the increase in features of multi-temporal image dataset helped increase performance for all the classifiers, the increase in sample size had a different effect on the classifier performance. While tree-based classifiers such as RF and CART show a slight increase in overall accuracy with sample size, kernel-based SVM and RVM classifiers showed better overall accuracy in smaller sample sizes. This is due the underlying behaviour of SVM and RVM which generalizes well by selecting a smaller subset of useful vectors for training, no matter the size of training sample. Similar results have been obtained in the study of Pal & Foody (2012) where RVM and SVM performed well for smaller training sets of ETM+ data. In fact, the authors even suggest RVM can be an alternative to SVM at smaller sample sizes as it uses much lesser training vectors than SVM. This can also be seen in Figure 5-6 which shows a much larger decrease in accuracy for RVM within smaller sample size of 1575 and 2250 pixels, indicating probability of better performance of RVM in sample sizes lesser than 1575. The higher accuracy results achieved for smaller

sample size and increased dimensionality during multi-class classification by SVM is similar to the multi-class image classification results achieved by Pal & Mather (2005) where a small sample set of 2700 pixels for 7 crop types gave an accuracy of 87%. Though few literatures such as Mountrakis et al. (2011) speak of less sensitivity of SVM to sample size, the current study showed decrease in accuracy with sample size. This is an indication that chosen input parameter value may not be most suitable. SVM's are highly sensitive to input parameters and the choice of C and kernel could bring this effect on the classifier. Since the study contains large dataset, linear kernel was used on a non-linear dataset. Further tuning of parameter C might help achieve better results for SVM in higher sample size.

On the contrary to RVM and SVM, CART and RF classifier showed small increase in overall accuracy with sample size (Deng & Wu, 2013). But the increase in performance has been small and can be noticed that these classifiers were comparatively least affected by sample size. Irrespective of sample size, RF performed distinctly well compared other classifiers, followed by CART. The Z-test results between RF and all other classifiers in Table 5-10 indicated that RF outperformed significantly better than other classifier, with 5% probability that the successful performance could be by chance. While some literatures report better performance of SVM over CART (Shao & Lunetta, 2012), few literatures have reported better performance of CART over SVM (Goldblatt et al., 2016). Such results can also be observed in this study where close association can be noticed between producer and use accuracies of CART and SVM. Additionally, the Z-statistical test of the pairwise performance of CART and SVM proved the comparative results to be insignificant. These results indicate that both the classifiers perform similar and can be used as an alternative to one other. In such cases, as suggested by Congalton & Green (2010) it is best to choose lesser complex and faster algorithm of the two for classification, CART. The better performance of CART over SVM could be favoured due to the quality of training samples.

As discussed in Section 6.1, quality of training sample affects the classifier performance and current study contains certain unavoidable errors in sample due to large study area and unavailability of highly accurate reference map. Due to this certain classes such as Shrubland and Grassland report a producer and user accuracy of less than 50% for all classifiers. As observed in and Figure 6-2 spectral profile graphs, Shrubland and Grassland overlaps with most of the other classes. This indicates impure or mislabelled training samples for these two classes. But the results for shrubland and grassland in Figure 5-4 and Figure 5-5 indicate different levels of sensitivity of classifiers to training data quality. SVM and RVM perform relatively low and are more sensitive compared to RF and CART. Foody et al. (2016) made a similar observation with SVM where the accuracy of the classifier decreased when intentionally mislabelled training data was introduced. In such cases, adding more features into the dataset will help SVM and RVM perform better in identifying classes mislabelled classes. The affinity to boundary pixels and ant-boundary nature of SVM and RVM respectively makes more sensitive to training data quality. Other Studies on RF classifiers, such as Mellor, Boukir, Haywood, & Jones (2015) have also reported low sensitivity of RF classifiers to training data quality. Certain other classes such as cropland show a low performance due to their spectral similarity with fallow land during dry seasons ((c)). This can be further affirmed by visual interpretation, which revealed good training data for Cropland, but due to the variations in multi-temporal data and the presence of dry croplands in some seasons, the overlap between fallow land and dry cropland can be reasoned out. Inclusion of additional features such as bare land index might help improve the accuracy of Cropland class. Even in this scenario, RF identified and separated classes more successfully than other classifiers. Irrespective of the situation, RF has proved to outperform other machine learning classifiers such as CART, SVM and RVM.

Classification studies on larger study areas is often needed and image processing tasks on such scale is a complex task and requires heavy computational environment. GEE made the processing of multi-temporal landsat-8 image series and classification on large study areas an easy task, which would otherwise require

intensive computation handling hardware. Additionally, GEE provides flexibility to perform most of the image processing tasks by providing various methods and datasets. GEE provides various machine learning classifiers such as RF, CART and SVM. Due to the unavailability of RVM in GEE, an external implementation of RVM was integrated with GEE to perform all the intensive tasks. The only drawback of GEE w.r.t the current research was the unavailability of tools to perform geo-statistical sampling processes such as semi-variance calculations and simulated sampling.

7. CONCLUSIONS AND FUTURE RECOMMENDATIONS

The section **7.1** concludes the findings of this research by answering the research questions proposed in **Chapter 1**. Future Scope of this work are included in **Chapter 7.2**

7.1. Conclusions

LULC maps are important from various aspects and extracting this information from remotely sensed images has been an area of interest for many decades. Among the many image classification techniques present, in recent year machine learning classifiers have been reported to produce highly accurate classification results. Accuracies of these classifiers are impacted by various factors and an understanding of such factors help improve the classification results. This research mainly aims to analyse different machine learning classifiers such as RF, CART, SVM and RVM on GEE, under the influence of certain factors such as sample size, data dimension and provide a comparative analysis of their sensitivity to other factors. The research also focuses particularly on sampling techniques for training data which requires more attention according to literatures. Three different stratified sampling techniques were studied among which SRS(Prop) method produced highest overall accuracy but under-represented smaller classes. SRS(Eq) gave a comparable overall accuracy and produced a balanced training dataset mapping even the smaller classes with good accuracy. While these methods produced a random distribution of samples, SSS method generated an even spread of training samples by considering the intra-class variability in terms of semi-variance and Spatial Simulated using MMSD. The fixed number of heterogeneous pixels produced good class-level accuracies. With all three methods favoured at different times, choice of sampling technique to obtain training samples depends on the requirement of a particular study. The effect of sampling can also be associated with the accuracy of reference maps to certain extent. With the absence of ground truth data in most of the studies, dependency on reference maps are higher on large areas. Hence there is a need for more freely accessible high resolution LULC maps which are regularly updated.

Accuracies of LULC maps also depend on the classifiers. Upon comparison we observed that RF and CART performed relatively well in different sample sizes, while SVM and RVM showed a decrease of 6.09% and 8.87% respectively in performance. SVM and RVM reported strongest performance in smaller sample sizes, which makes them an attractive classifier when training samples are limited. RVM substantially decreased the required training sample size to around 2 pixels/class by choosing only the relevant vectors. Additionally, the probabilistic output of RVM gives information about the uncertainty in class allocation and helps identify the sources of error. A multi-temporal data is best classified if the dataset includes the variation during the study period. With such high dimensional dataset of Landsat-8 consisting of 60 features, RVM and RF made maximum use of the additional data with 8% increase in overall accuracy while SVM and CART showed relatively less improvement in classification results. The reference map used in the study contains certain errors and the presence of misclassified training pixels helped observe the sensitivity of classifiers. Tree-based RF and CART classifier were very less sensitive to such samples while the kernel-based SVM and RVM showed high sensitive to the quality of training samples.

This study concludes that RF, CART, SVM and RVM are all powerful classifiers for LULC classifications. RVM, a Bayesian probabilistic classifier shows great potential with smaller training samples and must be explored further for LULC classifications of smaller areas by comparing its performance with other machine learning classifiers. Our results also indicate that the choice of classifier depends on the study area, thematic accuracy, quality of training samples and requirement of the map. However, among all the classifiers, RF proves to be a more stable and reliable classifier with the results highly coherent with other literatures. This

research was majorly performed on Google Earth Engine, considering the computational intensity involved and data requirements of the study. This platform will be highly beneficial for national/global scale classification using multi-dimensional products.

Based on the results of the study, the proposed research questions are answered below:

1. What is the effect of different training sampling techniques on the accuracy of LULC classification?
   The study showed the change in overall, producer and user accuracies with the change in sampling techniques. The accuracy depends on the size of sample in each class, size of each class, number of classes and the way samples are collected. SRS(Eq) method should be preferred if aim is to achieve good accuracies at individual class levels, irrespective of their size. SRS(Prop) on the other hand should be preferred if aim is to obtain good overall accuracy. SSS sampling can be used in case of large intra-class variability and automatic determination of sample size is required.

   a. Do the sampling methods equally affect smaller and larger sized classes?
   Effect of sampling method on each class depends on its size. SRS(Prop) method assigns less samples for smaller classes due to which classifier may not effectively map rare classes. SRS(Eq) method is assigns equal number of samples to all classes and is more beneficiary for smaller classes.

   b. What are the advantages and disadvantages of different methods?
   Described in Table 6-1.

2. How does RVM perform in classifying different LULC classes? What is the effect of training sample size on the classification result?

   RVM gave an overall accuracy of 64.01% for a small sample size of 175 pixels/class. It does so with very small number of relevant vectors. The classifier can even handle high dimensional dataset to obtain an improved classification result. The probabilistic output of RVM for each case, helped analyse the problematic classes and to verify the presence of misclassified training pixels. Also RVM requires good quality of training samples to discriminate the spectrally similar classes such as Fallow Land and Dry River Bed. Accuracy of RVM classifier decreases with increase in sample size. RVM is at its best at very small samples sizes.

3. How well do the in-built machine learning methods of GEE such as Random Forest, SVM and CART perform on multi-temporal satellite images in discriminating land cover classes of interest?
   For a given sample size of 1575 pixels, RF could discriminate a large study area of 9 classes with 79.58% accuracy. CART being the second best performer performed with an overall accuracy of 72.08% while SVM gave an accuracy of 69.63%.

   a. Which is the overall best performing classifier
   In all sample sizes and datasets, RF performed better than other classifiers with more than 95% confidence. RF was also least sensitive to training sample quality which is a great advantage in remote sensing field where the training samples are always prone to errors.

   b. How well do the classifiers perform with respect to each other?
   RF gives higher overall accuracies, user and producer accuracies compared to CART and SVM. While RF and CART show an increase in performance in samples size, SVM shows negative effect. Though CART shows better results than SVM, the statistical tests prove that CART and SVM perform similar, and can be used as an alternative to each other.

c. To what extent does the integrated RVM classifier perform compared to RF, CART and SVM?
For smaller training samples, RVM produces comparable results to SVM in terms of overall accuracies. On the other hand, RF performed 15.5% better than RVM while CART performed 8.05% better. Unlike CART and RF, RVM showed high sensitivity to sample size and quality.

7.2. Future Recommendations

7.2.1. Systematic Sampling using SSA

- Restrict the perturbation of samples during SSA to the range of the class. This will help further reduce the complexity of SSA where perturbation to large distances will not add any value in computation of MMSD.
- Compare study of systematic sampling efficiency in terms of computation and accuracy when using an initial sample distribution of Random samples and ones created using Range values.

7.2.2. Sampling for training data

- Based on the type and area of the class, apply different sampling techniques to different classes. Evaluate the training sample quality and LULC classifier performance on such data.

7.2.3. Machine Learning Classifiers

- Understand the relation between thematic accuracy and size of study area for RVM.
- Applicability of probabilistic output of RVM in quantifying the quality of training samples.
- Effect of fusing multi-temporal Sentinel-2 and Landsat-8 data to obtain high resolution LULC map at national level using red-edge bands of Sentinel-2 and thermal bands of Landsat-8 with RF classifier
- Improving training data quality obtained from a reference map

LIST OF REFERENCES

Abdulkareem, J. H., Sulaiman, W. N. A., Pradhan, B., & Jamil, N. R. (2018). Relationship between design floods and land use land cover (LULC) changes in a tropical complex catchment. *Arabian Journal of Geosciences*, *11*(14), 376. https://doi.org/10.1007/s12517-018-3702-4

Aboelnour, M., & Engel, B. A. (2018). Application of Remote Sensing Techniques and Geographic Information Systems to Analyze Land Surface Temperature in Response to Land Use/Land Cover Change in Greater Cairo Region, Egypt. *Journal of Geographic Information System*, *10*(01), 57–88. https://doi.org/10.4236/jgis.2018.101003

Aguilar, R., Zurita-Milla, R., Izquierdo-Verdiguier, E., & de By, R. A. (2018). A cloud-based multi-temporal ensemble classifier to map smallholder farming systems. *Remote Sensing*, *10*(5), 729. https://doi.org/10.3390/rs10050729

Aksoy, S., Koperski, K., Tusk, C., Marchisio, G., & Tilton, J. C. (2005). Learning bayesian classifiers for scene classification with a visual grammar. *IEEE Transactions on Geoscience and Remote Sensing*, *43*(3), 581–589. https://doi.org/10.1109/TGRS.2004.839547

Alonso, M., Malpica, J., & Agirre, A. de. (2011). Consequences of the Hughes phenomenon on some classification Techniques. *In ASPRS 2011 Annual Conference*, (May), 1–5. Retrieved from http://www.asprs.org/wp-content/uploads/2010/12/Alonso.pdf%5Cnhttp://info.asprs.org/publications/proceedings/Milwaukee2011/files/Alonso.pdf

Azzari, G., & Lobell, D. B. (2017). Landsat-based classification in the cloud: An opportunity for a paradigm shift in land cover monitoring. *Remote Sensing of Environment*, *202*, 64–74. https://doi.org/10.1016/j.rse.2017.05.025

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Beuchle, R., Grecchi, R. C., Shimabukuro, Y. E., Seliger, R., Eva, H. D., Sano, E., & Achard, F. (2015). Land cover changes in the Brazilian Cerrado and Caatinga biomes from 1990 to 2010 based on a systematic remote sensing sampling approach. *Applied Geography*, *58*, 116–127. https://doi.org/10.1016/J.APGEOG.2015.01.017

Bicheron, P., Defourny, P., Brockmann, C., Schouten, L., Vancutsem, C., Huc, M., … Herold, M. (2008). GlobCover 2005–Products description and validation report, Version 2.1, 2008. *Available on the ESA IONIA Website (Http://Ionia1. Esrin. Esa. Int/)*.

Bittencourt, H. R., & Clarke, R. T. (2003). Use of classification and regression trees (CART) to classify remotely-sensed digital images. In *Geoscience and Remote Sensing Symposium, 2003. IGARSS'03. Proceedings. 2003 IEEE International* (Vol. 6, pp. 3751–3753). IEEE. https://doi.org/10.1109/IGARSS.2003.1295258

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classiiers. In *5th Annual Workshop on Com- putational Learning Theory (COLT '92)* (pp. 144–152). New York, USA. Retrieved from http://www.svms.org/training/BOGV92.pdf

Boyd, D. S., Sanchez-Hernandez, C., & Foody, G. M. (2006). Mapping a specific class for priority habitats monitoring from satellite sensor data. *International Journal of Remote Sensing*, *27*(13), 2631–2644. https://doi.org/10.1080/01431160600554348

Breiman, L. (1996). *Bagging Predictors* (Vol. 24). Berkeley: Kluwer Academic Publishers. Retrieved from https://link.springer.com/content/pdf/10.1023%2FA%3A1018054314350.pdf

Breiman, L. (2001). *Random Forests* (Vol. 45). Berkeley. Retrieved from https://link.springer.com/content/pdf/10.1023%2FA%3A1010933404324.pdf

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification And Regression Trees*. Boca Raton: Chapman & Hall/CRC.

Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R. (2015). Improving kNN multi-label classification in Prototype Selection scenarios using class proposals. *Pattern Recognition*, *48*(5), 1608–1622. https://doi.org/10.1016/J.PATCOG.2014.11.015

Chapelle, O., & Bousquet, O. (2002). *Choosing Multiple Parameters for Support Vector Machines* (Vol. 46). Retrieved from https://link.springer.com/content/pdf/10.1023%2FA%3A1012450327387.pdf

Chen, B., Pan, Y., Wang, J., Fu, Z., Zeng, Z., Zhou, Y., & Zhang, Y. (2013). Even sampling designs generation by efficient spatial simulated annealing. *Mathematical and Computer Modelling*, *58*(3–4), 670–676. https://doi.org/10.1016/j.mcm.2011.10.035

Cochran, W. G. (1953). *Sampling Techniques*. https://doi.org/10.2307/1268167

Congalton, R. G., & Green, K. (2010). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices* (Vol. 25). Boca Raton: CRS Press. https://doi.org/10.1111/j.1477-9730.2010.00574_2.x

Cortes, C., Vapnik, V., & Saitta, L. (1995). *Support-Vector Networks. Machine Learning* (Vol. 20). Kluwer Academic Publishers. Retrieved from https://link.springer.com/content/pdf/10.1007%2FBF00994018.pdf

Deng, C., & Wu, C. (2013). The use of single-date MODIS imagery for estimating large-scale urban impervious surface fraction with spectral mixture analysis and machine learning techniques. *ISPRS Journal of Photogrammetry and Remote Sensing*, *86*, 100–110. https://doi.org/10.1016/J.ISPRSJPRS.2013.09.010

Dong, J., Xiao, X., Menarguez, M. A., Zhang, G., Qin, Y., Thau, D., … Moore, B. (2016). Mapping paddy rice planting area in northeastern Asia with Landsat 8 images, phenology-based algorithm and Google Earth Engine. *Remote Sensing of Environment*, *185*, 142–154. https://doi.org/10.1016/J.RSE.2016.02.016

Du, P., Samat, A., Waske, B., Liu, S., & Li, Z. (2015). Random Forest and Rotation Forest for fully polarized SAR image classification using polarimetric and spatial features. *ISPRS Journal of Photogrammetry and Remote Sensing*, *105*, 38–53. https://doi.org/10.1016/J.ISPRSJPRS.2015.03.002

Feller, W. (1971). *An Introduction To Probability Theory And Its Applications . Vol. II.* (Second). New York, USA: Willey.

Foody, G. M. (2008). RVM-based multi-class classification of remotely sensed data. *International Journal of Remote Sensing*, *29*(6), 1817–1823. https://doi.org/10.1080/01431160701822115org/10.1080/01431160701822115

Foody, G. M. (2009). Sample size determination for image classification accuracy assessment and comparison. *International Journal of Remote Sensing*, *30*, 5273–5291. https://doi.org/10.1080/01431160903130937org/10.1080/01431160903130937

Foody, G. M., & Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(6), 1335–1343. https://doi.org/10.1109/TGRS.2004.827257

Foody, G. M., Mathur, A., Sanchez-Hernandez, C., & Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*, *104*(1), 1–14. https://doi.org/10.1016/J.RSE.2006.03.004

Foody, G., Pal, M., Rocchini, D., Garzon-Lopez, C., Bastin, L., Foody, G. M., … Bastin, L. (2016). The Sensitivity of Mapping Methods to Reference Data Quality: Training Supervised Image Classifications with Imperfect Reference Data. *ISPRS International Journal of Geo-Information*, *5*(11), 199. https://doi.org/10.3390/ijgi5110199

Friedl, M. ., McIver, D. ., Hodges, J. C. ., Zhang, X. ., Muchoney, D., Strahler, A. ., … Schaaf, C. (2002). Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, *83*(1–2), 287–302. https://doi.org/10.1016/S0034-4257(02)00078-0

Gallego, F. J. (2005). Stratified sampling of satellite images with a systematic grid of points. *ISPRS Journal of Photogrammetry and Remote Sensing*, *59*(6), 369–376. https://doi.org/10.1016/J.ISPRSJPRS.2005.10.001

Ghimire, B., Rogan, J., Galiano, V. R., Panday, P., & Neeti, N. (2012). An Evaluation of Bagging, Boosting, and Random Forests for Land-Cover Classification in Cape Cod, Massachusetts, USA. *GIScience & Remote Sensing*, *49*(5), 623–643. https://doi.org/10.2747/1548-1603.49.5.623

Ghosh, A., Sharma, R., & Joshi, P. K. (2014). Random forest classification of urban landscape using Landsat archive and ancillary data: Combining seasonal maps with decision level fusion. *Applied Geography*, *48*, 31–41. https://doi.org/10.1016/j.apgeog.2014.01.003

Giri, C., Pengra, B., Long, J., & Loveland, T. R. (2013). Next generation of global land cover characterization, mapping, and monitoring. *International Journal of Applied Earth Observation and Geoinformation*, *25*, 30–37. https://doi.org/10.1016/J.JAG.2013.03.005

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters*, *27*(4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Goldblatt, R., You, W., Hanson, G., Khandelwal, A., Goldblatt, R., You, W., … Khandelwal, A. K. (2016). Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine. *Remote Sensing*, *8*(8), 634. https://doi.org/10.3390/rs8080634

Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., Liang, L., … Chen, J. (2013). Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data. *International Journal of Remote Sensing*, *34*(7), 2607–2654. https://doi.org/10.1080/01431161.2012.748992

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Guan, H., Li, J., Chapman, M., Deng, F., Ji, Z., & Yang, X. (2013). Integration of orthoimagery and lidar data for object-based urban thematic mapping using random forests. *International Journal of Remote Sensing*, *34*(14), 5166–5186. https://doi.org/10.1080/01431161.2013.788261

Guru, B., & Aravind, S. M. (2015). Land use land cover changes in pre- and postearthquake affected area using geoinformatics - Western Coast of Gujarat, India. *Disaster Advances*, *8*(4), 1–14.

Hansen, M. C., Potapov, P. V, Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., … Townshend, J. R. G. (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, *342*(6160), 850–853. https://doi.org/10.1126/science.1244693

Heydari, S. S., & Mountrakis, G. (2018). Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. *Remote Sensing of Environment*, *204*, 648–658. https://doi.org/10.1016/j.rse.2017.09.035

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). *A Practical Guide to Support Vector Classification*. https://doi.org/10.1007/s11119-014-9370-9

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, *23*(4), 725–749. https://doi.org/10.1080/01431160110040323

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, *42*(2), 513–529. https://doi.org/10.1109/TSMCB.2011.2168604

Jia, K., Liang, S., Wei, X., Yao, Y., Su, Y., Jiang, B., & Wang, X. (2014). Land Cover Classification of Landsat Data with Phenological Features Extracted from Time Series MODIS NDVI Data. *Remote Sensing*, *6*(11), 11518–11532. https://doi.org/10.3390/rs61111518

Jin, H., Stehman, S. V, & Mountrakis, G. (2014). Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado. *International Journal of Remote Sensing*, *35*(6), 2067–2081. https://doi.org/10.1080/01431161.2014.885152

Johnson, A., Truax, D. D., & O'Hara, C. G. (2002). Remote Sensing, GIS and Land Use and Land Cover Mapping along the I-10 corridor. In S. Morain & A. Budge (Eds.), *Integrating Remote Sensing at the Global, Regional and Local Scale.Proceedings of Pecora 15/Land Satellite Information IV Conference, American Society for Photogrammetry and Remote Sensing,Environmental Protection Agency, NASA, Department of Transportatio* (p. 9). Denver, USA. Retrieved from http://www.isprs.org/proceedings/XXXIV/part1/paper/00085.pdf

Khatami, R., Mountrakis, G., & Stehman, S. V. (2016). A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment*, *177*, 89–100. https://doi.org/10.1016/J.RSE.2016.02.028

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*, *14*(2), 1137–1145. Retrieved from http://robotics.stanford.edu/~ronnyk

Lam, N. S.-N. (2008). Methodologies for Mapping Land Cover/Land Use and its Change. In S. Liang (Ed.), *Advances in Land Remote Sensing* (pp. 341–367). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-6450-0_13

Lawrence, R. L., & Wrlght, A. (2001). Rule-Based Classification Systems Using Classification and Regression Tree (CART) Analysis. *Photogrammetric Engineering and Remote Sensing*, *67*(10), 1137–1142.

Retrieved from https://pdfs.semanticscholar.org/0a3c/ec5d1c1c1ba8ee2ce44dcae1b2bcc93c5519.pdf
Li, C., Wang, J., Wang, L., Hu, L., Gong, P., Li, C., … Gong, P. (2014). Comparison of Classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper Imagery. *Remote Sensing*, *6*(2), 964–983. https://doi.org/10.3390/rs6020964
Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, *164*, 324–333. https://doi.org/10.1016/j.rse.2015.04.021
Loveland, T. R., & Belward, A. S. (1997). The IGBP-DIS global 1km land cover data set, DISCover: First results. *International Journal of Remote Sensing*, *18*(15), 3289–3295. https://doi.org/10.1080/014311697217099org/10.1080/014311697217099
Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, *28*(5), 823–870. https://doi.org/10.1080/01431160600746456
Lück, W., & van Niekerk, A. (2016). Evaluation of a rule-based compositing technique for Landsat-5 TM and Landsat-7 ETM+ images. *International Journal of Applied Earth Observation and Geoinformation*, *47*, 1–14. https://doi.org/10.1016/J.JAG.2015.11.019
M.E.Tipping. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, *1*, 211–244. Retrieved from http://www.jmlr.org/papers/volume1/tipping01a/tipping01a.pdf
Mantero, P., Moser, G., Member, S., Serpico, S. B., & Member, S. (2005). Partially Supervised Classification of Remote Sensing Images Through SVM-Based Probability Density Estimation. *IEEE Transactions on Geoscience and Remote Sensing*, *43*(3), 559–570.
Marconcini, M., Camps-Valls, G., & Bruzzone, L. (2009). A Composite Semisupervised SVM for Classification of Hyperspectral Images. *IEEE Geoscience and Remote Sensing Letters*, *6*(2), 234–238. https://doi.org/10.1109/LGRS.2008.2009324
Mas, J. F., & Flores, J. J. (2008). The application of artificial neural networks to the analysis of remotely sensed data. *International Journal of Remote Sensing*, *29*(3), 617–663. https://doi.org/10.1080/01431160701352154
Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343
McBratney, A. B., Webster, R., & Burgess, T. M. (1981). The design of optimal sampling schemes for local estimation and mapping of of regionalized variables—I: Theory and method. *Computers & Geosciences*, *7*(4), 331–334. https://doi.org/10.1016/0098-3004(81)90077-7
Mellor, A., Boukir, S., Haywood, A., & Jones, S. (2015). Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. *ISPRS Journal of Photogrammetry and Remote Sensing*, *105*, 155–168. https://doi.org/10.1016/J.ISPRSJPRS.2015.03.014
Midekisa, A., Holl, F., Savory, D. J., Andrade-pacheco, R., Gething, W., Bennett, A., & Sturrock, H. J. W. (2017). Mapping land cover change over continental Africa using Landsat and Google Earth Engine cloud computing. *PloS One*, *12*(9), e0184926. https://doi.org/10.1371/journal.pone.0184926
Minasny, B., McBratney, A. B., & Walvoort, D. J. J. (2007). The variance quadtree algorithm: Use for spatial sampling design. *Computers & Geosciences*, *33*(3), 383–392. https://doi.org/10.1016/J.CAGEO.2006.08.009
Mochizuki, S., & Murakami, T. (2012). Accuracy Comparison Of Land Cover Mapping Using The Object-Oriented Image Classification With Machine Learning Algorithms. In *The 33rd Asian Conference on Remote Sensing*. Pattaya, Thailand. Retrieved from https://pdfs.semanticscholar.org/e161/8c75b6ac4271da2bf8a071c51b674a99f8e1.pdf
Montanari, R., Souza, G. S. A., Pereira, G. T., Marques, J., Siqueira, D. S., & Siqueira, G. M. (2012). The use of scaled semivariograms to plan soil sampling in sugarcane fields. *Precision Agriculture*, *13*(5), 542–552. https://doi.org/10.1007/s11119-012-9265-6
Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001
Nery, T., Sadler, R., Solis-Aulestia, M., White, B., Polyakov, M., & Chalak, M. (2016). Comparing

supervised algorithms in Land Use and Land Cover classification of a Landsat time-series. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 5165–5168). IEEE. https://doi.org/10.1109/IGARSS.2016.7730346

Padarian, J., Minasny, B., & McBratney, A. B. (2015). Using Google's cloud-based platform for digital soil mapping. *Computers & Geosciences*, *83*, 80–88. https://doi.org/10.1016/j.cageo.2015.06.023

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Pal, M., & Foody, G. M. (2012). Evaluation of SVM, RVM and SMLR for accurate image classification with limited ground data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*(5), 1344–1355. https://doi.org/10.1109/JSTARS.2012.2215310

Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, *86*(4), 554–565. https://doi.org/10.1016/S0034-4257(03)00132-9

Pal, M., & Mather, P. M. (2004). Assessment of the effectiveness of support vector machines for hyperspectral data. *Future Generation Computer Systems*, *20*(7), 1215–1225. https://doi.org/10.1016/J.FUTURE.2003.11.011

Pal, M., & Mather, P. M. (2005). Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, *26*(5), 1007–1011. https://doi.org/10.1080/01431160512331314083

Patel, N. N., Angiuli, E., Gamba, P., Gaughan, A., Lisini, G., Stevens, F. R., … Trianni, G. (2015). Multitemporal settlement and population mapping from Landsat using Google Earth Engine. *International Journal of Applied Earth Observations and Geoinformation*, *35*, 199–208. https://doi.org/10.1016/j.jag.2014.09.005

Pielke, R. A., Pitman, A., Niyogi, D., Mahmood, R., McAlpine, C., Hossain, F., … de Noblet, N. (2011). Land use/land cover changes and climate: modeling analysis and observational evidence. *Wiley Interdisciplinary Reviews: Climate Change*, *2*(6), 828–850. https://doi.org/10.1002/wcc.144

Plourde, L., & Congalton, R. G. (2003). Sampling Method and Sample Placement: How Do They Affect the Accuracy of Remotely Sensed Maps. *Photogrammetric Engineering & Remote Sensing*, *69*(3), 289–297. https://doi.org/10.1080/1061186021000038391

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*, 93–104. https://doi.org/10.1016/J.ISPRSJPRS.2011.11.002

Roland Colditz, R. (2006). An Evaluation of Different Training Sample Allocation Schemes for Discrete and Continuous Land Cover Classification Using Decision Tree-Based Algorithms. *International Journal of Remote Sensing*, *7*, 7. https://doi.org/10.3390/rs70809655

Rossiter, D. G. (2014). *Statistical methods for accuracy assesment of classified thematic maps*. Retrieved from www.itc.nl/personal/rossiter

Roy, P. S., Kushwaha, S., Murthy, M., & Roy, A. (2012). *Biodiversity Characterisation at Landscape Level: National Assessment*.

Samuel-Rosa, A., Heuvelink, G., Vasques, G., & Anjos, L. (2015). spsann - optimization of sample patterns using spatial simulated annealing. *EGU General Assembly 2015, Held 12-17 April, 2015 in Vienna, Austria. Id.7780*, *17*. Retrieved from http://adsabs.harvard.edu/abs/2015EGUGA..17.7780S

Shalaby, A., & Tateishi, R. (2007). Remote sensing and GIS for mapping and monitoring land cover and land-use changes in the Northwestern coastal zone of Egypt. *Applied Geography*, *27*(1), 28–41. https://doi.org/10.1016/J.APGEOG.2006.09.004

Shao, Y., & Lunetta, R. S. (2012). 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*, *70*, 78–87. https://doi.org/10.1016/J.ISPRSJPRS.2012.04.001

Shaumyan, A. (2017). Python package for Bayesian Machine Learning with scikit-learn API. Retrieved from https://github.com/AmazaspShumik/sklearn-bayes

Shelestov, A., Lavreniuk, M., Kussul, N., Novikov, A., & Skakun, S. (2017). Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping. *Frontiers in Earth Science*, *5*, 17. https://doi.org/10.3389/feart.2017.00017

Srivastava, P. K., Han, D., Rico-Ramirez, M. A., Bray, M., & Islam, T. (2012). Selection of classification

techniques for land use/land cover change investigation. *Advances in Space Research*, *50*(9), 1250–1265. https://doi.org/10.1016/J.ASR.2012.06.032

Stehman, S. V. (1992). Comparison of Systematic and Random Sampling for Estimating the Accuracy of Maps Generated from Remotely Sensed Data. *Photogrammetric Engineering & Remote Sensing*, *58*(9), 1343–1350. Retrieved from https://www.asprs.org/wp-content/uploads/pers/1992journal/sep/1992_sep_1343-1350.pdf

Stehman, S. V. (2009). Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing*, *30*(20), 5243–5272. https://doi.org/10.1080/01431160903131000

Thunig, H., Wolf, N., Naumann, S., Siegmund, A., Jurgens, C., Uysal, C., & Maktav, D. (2011). Land use/land cover classification for applied urban planning - the challenge of automation. In *2011 Joint Urban Remote Sensing Event* (pp. 229–232). IEEE. https://doi.org/10.1109/JURSE.2011.5764762

Tipping, M. E. (2004). Bayesian Inference: An Introduction to Principles and Practice in Machine Learning. In *Advanced lectures on machine Learning* (pp. 41–62). Berlin, Heidelberg: Springer. Retrieved from http://www.miketipping.com/papers.htm

Tipping, M. E., & Faul, A. C. (2003). Fast Marginal Likelihood Maximisation for Sparse Bayesian Models. In C. M. Bishop & B. J. Frey (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Key West, FL. Retrieved from http://www.miketipping.com/papers.htm

Trianni, G., Angiuli, E., Lisini, G., & Gamba, P. (2014). Human settlements from Landsat data using Google Earth Engine. In *2014 IEEE Geoscience and Remote Sensing Symposium* (pp. 1473–1476). IEEE. https://doi.org/10.1109/IGARSS.2014.6946715

Tso, B., & Mather, P. M. (2009). *Classification Methods for Remotely Sensed Data* (Second). Boca Raton: CRC Press.

Tsoi, A. C., & Pearson, R. A. (1991). Comparison of three classification techniques, CART, C4.5 and Multi-Layer Perceptrons. In *Advances in neural information processing systems* (pp. 963–969). Retrieved from http://papers.nips.cc/paper/410-comparison-of-three-classification-techniques-cart-c45-and-multi-layer-perceptrons.pdf

Tumer, K., & Ghosh, J. (1996). Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition*, *29*(2), 341–348. https://doi.org/10.1016/0031-3203(95)00085-2

Tuteja, U. (2013). *Baseline Data on Horticultural Crops in Uttarakhand*. Retrieved from http://www.du.ac.in/du/uploads/Academics/centres_institutes/Agricultural_Eco/18.2013-Baseline horti_Uttarakhand Usha.pdf

Van Groenigen, J. W., & Stein, A. (1998). Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, *27*(5), 1078–1086.

Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10. https://doi.org/10.1016/J.GSF.2015.07.003

Waske, B., & Braun, M. (2009). Classifier ensembles for land cover mapping using multitemporal SAR imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *64*(5), 450–457. https://doi.org/10.1016/j.isprsjprs.2009.01.003

Xin Huang, & Liangpei Zhang. (2010). Comparison of Vector Stacking, Multi-SVMs Fuzzy Output, and Multi-SVMs Voting Methods for Multiscale VHR Urban Mapping. *IEEE Geoscience and Remote Sensing Letters*, *7*(2), 261–265. https://doi.org/10.1109/LGRS.2009.2032563

Yu, L., Liang, L., Wang, J., Zhao, Y., Cheng, Q., Hu, L., … Gong, P. (2014). Meta-discoveries from a synthesis of satellite-based land-cover mapping research. *International Journal of Remote Sensing*, *35*(13), 4573–4588. https://doi.org/10.1080/01431161.2014.930206

APPENDIX – A

1. Creation of combined Reference Map

The producer and user accuracies of Globcover (2015) and BCLL (2012) maps are shown in Table 7-1. River Bed class is not present in Globcover map for the given study are and hence the values are empty. Based on the results of producer and user accuracy, and visual interpretation, classes polygons were chosen from different maps as shown in Table *7-2*. River Bed, Water Bodies and Fallow Land were manually delineated as the study considered multi-temporal data and it was necessary to create polygons over regions which remains as Water Body, River Bed and Fallow Land throughout the year of 2017.

Table 7-1 : Producer and User Accuracies of Globcover and BCLL reference maps validated using test sample of 100 pixels/class [BU – Built-Up, CL – Cropland, FL – Fallow Land, EV – Evergreen Forest, DE – Deciduous Forest, SL – Shrubland, GL – Grassland, WB – Water Bodies, RB – River Bed].

Reference Map	Producer Accuracy (%)							
	BU	CL	EV	DE	SL	GL	WB	RB
BCLL	88.46	81.13	88.99	73.88	59.5	71.42	97.61	55.86
Globcover	93.26	85.18	89.89	57.49	48.39	25	97.75	
User Accuracy (%)								
	BU	CL	EV	DE	SL	GL	WB	RB
BCLL	94.84	86	97.97	99	72	15.30	41.84	81.81
Globcover	1	71.85	89	96396	15.30	0.02	1	-

Table 7-2: Sources of various classes in the final reference map

BCLL	Cropland	Evergreen Forest	Deciduous Forest	Shrubland	Grassland
Globcover	Built-Up				
Visual Interpretation	Fallow Land	Water Bodies	River Bed		

APPENDIX – B

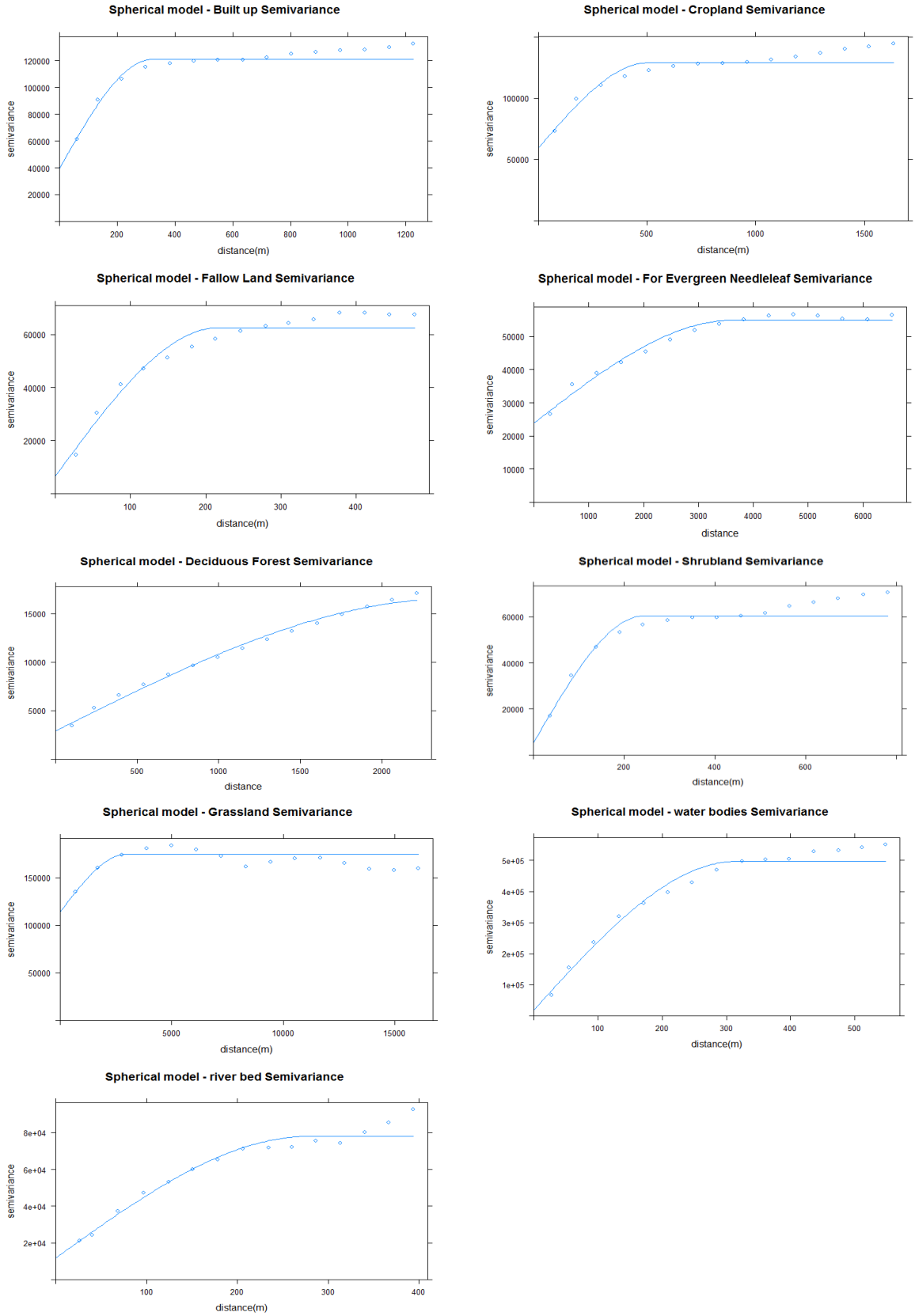


Figure 7-1: Spherical Model for semi-variogram of different classes

Table 7-3: Error Matrix for RF classification of 8 classes using stratified systematic samples generated at the distance of Range. Since sample size is less, small randomly generated sample of size 20/class is considered.

		REFERENCE MAP							
CLASSIFIED MAP		**BU**	**CL**	**EV**	**DE**	**SL**	**GL**	**WB**	**RB**
	BU	9	2	0	0	0	0	0	9
	CL	0	12	0	2	0	5	0	1
	EV	0	1	10	3	0	6	0	0
	DE	0	3	0	17	0	0	0	0
	SL	0	6	0	2	5	7	0	0
	GL	1	1	4	0	1	13	0	0
	WB	0	3	0	1	0	4	2	10
	RB	3	1	0	2	1	0	0	13
	Overall Accuracy: 36.65%								

APPENDIX –C

Table 7-4: Field visit data from few diverse regions of Dehradun District. The points were collected using Leica GPS device

Latitude	Longitude	Accuracy(m)	Class Identified
30*20'24.2998	77*59'36.8238	5	Deciduous Forest
30*20'37.0971	77*59'42.3578	4.3	Grassland
30*20'38.0590	78*00'00.8636	5	Grassland
30*20'24.5824	78*00'18.0817	3.1	Deciduous Forest
30*20'38.9706	77*56'42.9248	3.7	Shrubland
30*20'38.9302	77*56'19.7876	3.4	River Bed
30*20'41.8120	77*55'27.2510	1.8	Cropland
30*20'44.8369	77*55'30.4093	3.1	Cropland
30*20'49.4023	77*55'30.0374	2.6	Deciduous Forest
30*22'12.5140	77*49'59.0075	3.6	Built-Up
30*23'52.0472	77*48'23.2966	1.3	Fallow Land
30*23'50.5885	77*48'20.9264	1.6	River Bed
30'24'56.8242	77*49'11.7026	2.8	Cropland
30*24'54.8103	77*49'10.8062	2.9	Fallow Land
30*24'54.2941	77*49'11.9694	2	Shrubland
30*25'40.8860	77*49'31.8766	3.4	River Bed
30*26'28.8855	77*44'09.8318	2.2	Shrubland
30*26'30.7917	77*44'09.1255	1.85	Shrubland
30*26'33.3578	77*42'11.1384	2.5	Water Bodies
30*26'24.6094	77*40'20.0028	2.5	Water Bodies
30*26'09.2256	77*39'56.6820	2.2	Water Bodies
30*18'35.7696	77*00'28.0494	5.9	Built-Up

APPENDIX - D

This appendix shows results of classifiers while classifying using mean, media, standard 3-month composites of NDVI band as dataset (similar as D-2 dataset).

Table 7-5: Random Forest Classified Results for different tree and sample size, with NDVI band dataset similar to D-2

Parameters	No.of Pixels/Class		Accuracy(%)
No.of Trees	Training Samples	Validation Samples	NDVI
100	175	75	66.78
100	250	108	70.02
100	700	300	71.79
150	700	300	72.08
200	700	300	71.9
200	1400	600	70.36

Table 7-6: CART Classified Results for different tree and sample size, with NDVI band dataset similar to D-2

Parameters	No.of Pixels/Class		Accuracy(%)
Cross-Validation Factor	Training Samples	Validation Samples	NDVI
10	175	75	61.4
5	250	108	62.63
10	700	300	61.96
5	700	300	63.39
10	700	300	63.39
5	1400	600	65.34
10	1400	600	65.34

Table 7-7: SVM Classified Results for different tree and sample size, with NDVI band dataset similar to D-2

Parameters	No.of Pixels/Class		Accuracy(%)
Kernel Type/ Cost/ Gamma	Training Samples	Validation Samples	NDVI
Linear/3510	175	75	68.7
Linear/3510	250	108	67.96
RBF/10000/22	700	300	13
Linear/2048	700	300	54.3
Linear/3510	700	300	65.66
Linear/3510	1400	600	65.66