Mapping Tree Species Richness of Tropical Forest using Airborne Hyperspectral Remote Sensing

ANUSHREE BADOLA
March, 2019

SUPERVISORS:
Dr. Hitendra Padalia
Dr. Mariana Belgiu
Advisor: Mr. Prabhakar Alok Verma

Mapping Tree Species Richness of Tropical Forest using Airborne Hyperspectral Remote Sensing

ANUSHREE BADOLA
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:
Dr. Hitendra Padalia
Dr. Mariana Belgiu
Advisor: Mr. Prabhakar Alok Verma

THESIS ASSESSMENT BOARD:
Prof. Dr. Ir. A. Stein (Chair)
Dr. S. P.S. Kushwaha (External Examiner, Dean former (IIRS))

ABSTRACT

The information regarding tree species in tropical forest is of high priority for the effective forest management, conservation, utilization and policy development. The Remote sensing data can be effectively used to provide species level information. The procurement of this information became possible due to the availability of the hyperspectral data. This study made use of airborne imaging spectroscopy to map tropical tree species richness in Shimoga, Karnataka, India. Hyperspectral imagery with spectral variation ranging from wavelength 400 nm – 2,500 nm acquired by the Airborne Visible and Infrared Imaging Spectrometer – Next Generation (AVIRIS-NG) sensor on 1st January 2016, was analyzed to map the tree species of the tropical forest. A field survey was conducted to collect the tree location data from the study area. Data was collected from the 25 different plots which were laid by using conditioned Latin Hypercube Sampling (cLHS) technique.

To map forest tree species, selection of suitable classification approach is required. This study attempts to develop a methodology for classification of hyperspectral data for species level classification. Recently ensemble classifiers have gained the importance in the scientific community for the classification of data having high number of features. Random forest (RF) is a popular ensemble classifier in which performance is dependent on the strength and diversity of the individual base classifier in the ensemble. This study aims to increase diversity in the individual classifiers of the ensemble thus improving its overall accuracy. This was done by modifying RF by transforming the variables at each node to another space using Principal Component Analysis (PCA). The transformation at each tree node, improved the classification performance. This new method, PCA based Rotation Random Forest (RoRF) was validated by comparing it with RF and Support Vector Machine (SVM). RoRF has attained an overall tree species accuracy of 52.76% while SVM and RF has shown an overall accuracy of 41.21% and 40.34% respectively. The performance of SVM, RF and RoRF were evaluated using McNemar test. The performance of RoRF was found to be significantly different from the other two classifiers. The performance of RF and SVM do not differ significantly. Since the best performance was shown by RoRF classifier so with this classified map, species richness of the study area was calculated and compared with the species richness recorded in the study area. This study attempted to classify 20 tropical tree species of the study area including both rare and dominant species. Those species which were present in groups were classified better than the sparsely spread species. Also, the species richness was more in the region of drainage depressions rather than the areas where ridges were present. This study concludes that AVIRIS-NG data has the potential for mapping tree species of tropical forest. Also, the RF performance can be improved by improving diversity in individual classifier by transforming the data at each node into another subspace. For the future work it would be interesting to apply different feature extraction methods in place of PCA and compare their performance. Also, for more effective richness measurement ancillary data like LiDAR can be integrated with hyperspectral data.

Key words: Classification, Hyperspectral remote sensing, Tropical forest, Tree species richness; PCA; SVM; Rotation Random forest

ACKNOWLEDGEMENTS

I would like to thank my parents for supporting me in every possible way through my studies. This research work was possible only due to their endless love and support.

I want to convey my heartfelt thanks to all my supervisors for their enthusiastic guidance, meticulous suggestions and sound counselling. I'm grateful to my supervisor Dr Hitendra Padalia, for guiding and supporting me through this research, it was only due to his relentless efforts and encouragement that the field visit was possible. Dr Mariana Belgiu has been extremely kind during our discussion meets. She motivated me and made it very comfortable for me to ask questions. Mr Prabhakar Alok Verma was highly approachable. He listened to all my doubts patiently, realised the roadblocks and gave effective solutions to the problems. My supervisors have been the pillars of support for this research.

My sincere thanks to all the IIRS and ITC faculty for enlightening me with the knowledge throughout the course duration. My gratitude goes to my IIRS course coordinator Dr Sameer Saran, his support and caring nature towards the students always brings confidence in us. He was always there to listen to our problem and encourage us.

I would also like to thank the Karnataka Forest Department, for organizing a peaceful and cordial stay in Shimoga. Mr S. Chandrashekhar (WL DCFO) was wonderful in organising the field visits, and the hospitality and comfortable stay provided by Mr Anthony Mariyappa (DCFO) was quite praiseworthy. Range forest officers Mr Shivaraj Mathad and Mr Lingaraddi R Mankani has been very supportive to me during my entire field visit. I want to thank all the staff of Karnataka Forest department including forest watchers and labours, without whom this field visit was not possible especially Mr Anthony Rego who has assisted me during the field work.

It gives me great pleasure to acknowledge my friends whose help, and collective efforts have been reflected in the completion of this thesis. Their support and motivation was a blessing for me. They have been with me during all ups and downs and supported me beyond the infinity. A special thanks to Anurag sir, Utsav, Abhisek, Anjali and Yogendra for their endless support and motivation throughout this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Motivation

Human beings have been a key driver for change in the functionality of Earth systems. Anthropogenic activities put a lot of pressure on Earth causing ecosystem depletion and change in Earth's climate (Rockström et al., 2009). To set limits on these activities, Stockholm Resilience Centre (SRC) has proposed nine planetary boundaries for sustainable functioning of Earth systems. Beyond the threshold of these boundaries, the risk of adverse and inevitable change in ecosystem increases (Stockholm Resilience Centre, 2012). Loss of biosphere integrity also known as loss of biodiversity is among the four boundaries that have already crossed the threshold limit. It is identified as the core component because it has drastic effects on Earth systems (Jaramillo & Destouni, 2015). Biodiversity loss is directly related to human welfare. For example, the international demand for timber results in the reduction of forest cover hence causes loss of regional biodiversity which can increase the risk of floods (Millennium Ecosystem Assessment, 2005). Preventing biodiversity loss is the need of the hour, and hence, it is necessary to monitor forests to avoid biodiversity loss. Species diversity estimation is a proven technique to keep a check over biodiversity loss.

To measure species diversity, there are two significant factors to be considered, i.e. species richness and evenness (Purvis & Hector, 2000). Species richness is the total number of number species present in a particular ecological community or region while, species evenness can be defined as the relative abundance of individual species in a community (Wilson, 1993). Mapping species richness is essential from both management and scientific point of view. For better management, a forest official can estimate the economic yield of the forest, and prepare an effective and productive working plan and also for documenting forest inventory.

From a scientific perspective, regions having a more significant number of different species are specially targeted for biodiversity conservation (Magurran, 2004). A researcher can further identify endemic and keystone species which are the key driver of an ecosystem. Mapping of species richness will provide a future scope for biodiversity loss estimation by studying temporal data to estimate the change detection. It is necessary because tropical forests are already under high risk due to encroachment (Bhat, Chandran, & Ramachandra, 2012) and illegal felling, causing forest fragmentation and increasing global warming (Clark, Roberts, & Clark, 2005). Tree species mapping not only supports management and scientific aspects but it has substantial social significance as well.

Species mapping adds support to Sustainable Development Goal (SDG) number 15, Life on Land, focusing on the target number 15.5, which aims to take suitable action to check natural habitat and biodiversity loss by 2020. According to the report of UNDP (2016), only 1% of about 80,000 tree species have been potentially used. Also, around 80% of the rural inhabitants around the world make use of direct plant extract for medicinal use which creates a need to keep a check on the availability and abundance of these species. Hence, this gives motivation to develop and improve techniques for mapping the different

tree species. With this motivation, this study attempts to map the species present in a tropical forest region in Western Ghats of India.

A major work on the forest type classification system of Indian forests was done by Champion & Seth (1968). They majorly classified the forest till the level of vegetation groups based on the climatic, edaphic and local conditions. Another major study on the classification and mapping of tree species of Indian forests was done by Roy et al. (2015) by using the IRS LISS-III multispectral data. In this study, they attempt to map the 17 gregarious tree species (tree species present in pure associations) like *Shore sp.*, *Tectona sp.*, *Bamboo sp.*, *Pinus sp.* etc. However, detailed species-level information is still missing which can be possible to achieve by using hyperspectral data.

These assessments draw focus towards mapping of species richness. Due to the vastness and remoteness of the forest, this analysis can always be better with remote sensing. Also, in a survey conducted by Felbermeier et al. (2010) among professionals of forestry, it has been seen that two third among them reported a deficiency in the information of forest and 90% of them supported the fact that application of remote sensing is the best way to bring improvements. Tree species classification by remote sensing has provided a broad application in assessment and monitoring of biodiversity, mapping of wildlife habitat, insect abundance, management of hazard and stress, invasive species mapping and conservation and sustainable management of the forest. Information regarding tree species is essential in understanding the ecology of the tree communities and the contribution of tree species to ecosystem services and functions (Fassnacht et al., 2016).

1.2. Background on mapping tree species by remote sensing

Over past two decades, studies on tree species mapping using remote sensing have increased exponentially. One reason for this exponential increase, is the availability of the hyperspectral and LIDAR data (Fassnacht et al., 2016). There have been many studies dedicated to remote sensing of plant diversity till date that have been following two extensive approaches; namely direct and indirect (Carleer & Wolff, 2004; M. Foody, Atkinson, Gething, Ravenhill, & Kelly, 2005; Turner et al., 2003). Direct approaches were based on the identification of species from the remotely sensed data and map them directly (Turner et al., 2003). Some of these studies involved identification of individual species, weeds and mangrove species. Indirect approaches are based on modelling the species distribution and diversity indices distribution like Fisher's alpha, Shannon diversity index (SDI), Simpson index (Akbari & Kalbi, 2017). Few studies also focussed on spectral heterogeneity aspect of the data, i.e. Spectral Variation Hypothesis (SVH) to assess the plant diversity or the habitat heterogeneity (Palmer et al., 2002; D. Rocchini et al., 2009). However, these traditional methods of image analysis fail to provide detailed individual species classification and hence do not fully make use of the available data (Foody & Cutler, 2003). Some studies are focussed on using Normalized Difference Vegetation Index (NDVI) (Madonsela et al., 2018; Pouteau, Gillespie, & Birnbaum, 2018). Wang et al. (2004) integrated the pixel-based approach and object-oriented classification for mapping different groups of mangrove species available on the Caribbean coast of Panama. Some of the researches were also focussed on the different types of spectral unmixing algorithms available and used them for mapping individual species of plants (Parker Williams & Hunt, 2002; Robichaud et al., 2007). Sobhan (2007) utilised a spectral unmixing technique for airborne hyperspectral imagery to detect shrubs and tree species composition at the pixel level using the HyMap image. But the research had a limitation of not being able to find the end member spectra which truly represented the species.

In recent years, many studies have been done on pixel-based classifiers, in which each pixel is independently processed, and its spectral information is taken as the input to the classifier. Support vector machine (SVM) is a kernel based classifier and has proved to perform better in case of hyperspectral data since it gives good results even with less training data (Bahria, Essoussi, & Limam, 2011). Also, multiple classifier systems (MCSs) or ensemble classifiers have gained importance in recent times, in which a pixel is classified based on the set of individual classifiers. MCSs make consideration of the diverse information received by individual ensemble classifiers which tend to increase the overall performance (Xia et.al., 2017). Random Forest (RF) has been a popular ensemble classifier used in various studies and has been increasingly applied for hyperspectral data for tree species richness (Ferreira et al., 2016). RF randomly selects a certain amount of variables at each node of each tree and then chooses the best splits out of those selected variables (Belgiu & Drăguţ, 2016). However, when individual tree data consist of too many unimportant variables, their usage tends to generate noise which leads to reduced ensemble performance. Hence, in recent years, modifying RF has gained interest in the research community. This study focuses on one of the modified version of RF that is Rotation Random Forest (RoRF) which incorporates Principal Component Analysis (PCA) to the random forest at each node (Rodriguez, Kuncheva, & Alonso, 2006). Then the result of RoRF is compared with both kernel-based method, i.e. SVM and the ensemble classifier, i.e. RF.

1.3. Problem statement

As discussed in the previous sections, classification studies for predicting species-level information of the tropical forest are still unable to give detailed results. Hence, it provides a scope of exploring newer classification techniques to map the species richness of the tropical forest by using hyperspectral data. This study aims to partially fulfil that gap by experimenting an unexplored classifier, RoRF.

1.4. Research identification

The objective of the research is to map the individual tree species of the tropical forests of Shimoga region of Karnataka in India to get the species richness map.

1.5. Research objectives

1. To implement RoRF classifier which relies on PCA to transform variables at each tree node to another space for classifying hyperspectral data.
2. To compare classification results obtained by the implemented RoRF classifier with those obtained by RF and SVM.
3. To create species richness map of the study area from the classified image.
4. To compare species richness obtained from the classified image with the one calculated from the field data.

1.6. Research questions

Research objective 1 and 2:

1. How does the PCA-based RoRF classifier perform in comparison with RF and SVM?
2. How to deal with the unclassified tree species classes that are present in the study area?

Research objective 3 and 4:

1. To what extent can we map tree species from the used hyperspectral data?
2. To what extent the species richness obtained from classified image differs from the species richness obtained from field data?

1.7. Innovation

In this study, hyperspectral data has been explored for tree species level classification. Classification of hyperspectral data has been carried out mostly by Maximum Likelihood Classifier (MLC) (Jia & Richards, 1994), SVM (Vapnik, 1995) and RF (Breiman, 2001). MLC can effectively classify low to moderate resolution data, but for high spectral resolution data, it fails due to the presence of high variance. SVM has proven to give good results for hyperspectral data even with the less training data, but it has the limitation of manually selecting the penalty parameter (C) and gamma (γ) (Burges, 1998). These parameters are further discussed in section 4.4.1.

RF is a type of ensemble classifier. It is a simple method since it requires an adjustment of only the number of trees and the number of input features used to split at each tree node (Breiman, 2001). RF gained popularity to classify hyperspectral data because it can efficiently deal with a high number of input features which are the total number of bands in this case. Due to high dimensionality (Hughes phenomenon, discussed in section 2.1) of the hyperspectral data, there are chances of having very high unimportant features. In such cases, RF sometimes have tendency to show reduced performance. This is because, for high accuracy, ensemble classifiers need to have high diversity in their base classifier which can be achieved if each partition of feature set can accommodate within the base classifiers with equal probability. But since RF creates the feature subsets by random selection, it reduces the probability of all possible feature subsets being different. Hence there is a requirement to introduce extra randomization for the ensemble. It can be implemented by applying PCA on a bootstrap sample which results in RoRF technique which also solves the problem of the curse of dimensionality (Rodriguez et al., 2006).

In this MSc Thesis, RoRF will be used for tree species mapping from hyperspectral satellite images. This classifier relies on the transformation of variables at each node to another space. According to Zhang & Suganthan (2014), this approach increases the diversity and accuracy of each decision tree built in the classifier. PCA is used to transform the data at each node. We investigate whether the transformation at each tree node increases the classification results. This research will be focussed on making use of PCA-based RoRF for the purpose of classification of hyperspectral imagery to identify different species in the selected study area. This research not only experiments the integrated RF- PCA approach but also fulfils the need for mapping of the species richness data for the selected region which has not been done yet.

1.8. Thesis outline

This thesis has been organised into six chapters. Chapter 1 introduces the motivation of the study along with the background of the topic, problem statement, research identification, objectives and research questions. Chapter 2 briefs about the literature related to the research. Chapter 3 provides a description of the study area, dataset and the software that were used in this research. Chapter 4 describes the method adopted to achieve the research objectives. Chapter 5 shows the results obtained after applying the method. Chapter 6 contains a discussion on the obtained results. Chapter 7 concludes the research provided with the answers to the research questions and future recommendations.

2. LITERATURE REVIEW

The following sections provide an insight of scientific literature related to the estimation of tree species richness in case of tropical forest making use of different multispectral and hyperspectral datasets and by using various classification techniques. Fassnacht et al. (2016) provided a broad analysis on the classification of different tree species using remotely sensed data in tropical forests. Maity et al. (2017) currently performing similar work related to species identification in the same study area as this research, is using the same sensor data, but the approach they considered for the classification and identification is different. They made use of absorption peak decomposition method for identification of nineteen species although they have been successful in identifying only three species till now. In the past, it has been proven that to accommodate complex, non-normal and multimodal within-class variations, RF and SVM are better-classifying techniques (Baldeck et al., 2015). Hence, the focus of this research is focussed on the above-stated classification techniques. The sections below are subcategorized into following different parts which explain in brief the above mentioned different topics:

- Hyperspectral remote sensing
- Mapping tropical forest tree species
- Methods used for tree species classification
- Comparative performance of classification algorithms

2.1. Hyperspectral remote sensing

So far, medium to high-resolution multispectral data such as Quickbird data (Rocchini et al., 2009) and Landsat data (Bhat, Chandran, & Ramachandra, 2012) have been used for classifying different types of vegetation. But the hyperspectral image analysis gained importance due to its ability to provide detailed contiguous spectral signature curves which cannot be obtained from multispectral data. Hyperspectral data has a high spectral resolution which enhances its capability of differentiating between different ground objects like vegetation, soil, minerals and rocks. Its narrow bandwidth gives detailed information of earth surface which is difficult to obtain from coarser bandwidth data like that of the multispectral sensor. However, due to the huge volume of the data with increased dimensionality, it has been a challenge to extract thematic information from the hyperspectral data. High dimensionality in the data, i.e. a large number of bands is of great advantage to classify more number of classes. But the size of training sample required to train the classifier is dependent on the number of bands in the dataset. It has been mentioned by Mather & Koch (2011) that a training data of size $10\,N$ to $30\,N$ per class is required to train the classifier, where N is the number of bands present in the dataset. Hence, in case of high dimensional dataset, the number of bands are very high resulting into requirement of the large training data per class to train the classifier. Therefore, the boon of the hyperspectral data i.e. large number of bands turns into the curse of dimensionality which is also known as Hughes Phenomenon which means for training the classifier; sample size increases exponentially with the total number of bands present in the data (Chutia et al., 2016). This "curse" can overcome by using the suitable classification algorithms discussed in section 2.3.

Hyperspectral imagery classification has begun since the 1980s (Goetz et al., 1985) by using traditional multispectral classification approaches, but they produced inconsistent classification results. Significant changes in this field have happened since 2004 with the development of advance classifiers like SVM, RF etc. However, in recent times, many developments have been done for improving advance classifiers to apply on hyperspectral data. Some of these advancements include RFs classification and regression methods, Random subspace ensembles and Rotation forest etc. (Chutia et al., 2016).

Hyperspectral data has proven to give better accuracy than active sensors like SAR and LiDAR especially in the identification of tree species in areas having rich biodiversity (Fassnacht et al., 2016). Studies have been conducted to map tree species with HYDICE sensor data (Clark et al., 2005), Hyperion sensor data (Kalacska et al., 2007) and AVIRIS sensor data (Carlson et al., 2007). In the case of Hyperion data, it was difficult to identify species richness of rich diversity areas. Also, its poor spatial resolution and cloud interference affected the quality of images. The different spectral patterns obtained with AVIRIS data gave better results than EO-1, Hyperion due to very high signal-to-noise ratio which means data contains more information and less noise (Carlson et al., 2007). Hyperspectral sensor data give high precision in species-level forest classification, species identification, canopy density, etc. (Wang & Zhao, 2016). George et al. (2014) conducted a study for tree species discrimination in the western Himalayan region in India. They utilised Hyperion data and had shown that it performs comparatively better than Landsat data. It is because it can capture the spectral variability of the plant species, which hence improves the forest tree species mapping.

2.2. Mapping tropical forest tree species

Many studies have been done to map the tree species in the tropical region. In the tropical forest, the wide variety of species makes it a challenging task to identify and map all the species. Even if training data is available for those species, classification output will not be very appealing since, with increase in the number of classes, accuracy tends to decrease. To prove this, Feret & Asner (2013) conducted an experimental study with 17 species and 50 samples per species tested over seven different classifiers and all of them resulted in a linearly decreasing trend of accuracy with respect to number of species. The result varied from approx. 85-95% accuracy with 2 classes and 25% to 75% accuracy with 17 classes.

In a study done over Costa Rica forest, seven different forest tree species were classified using HYDICE sensor data achieving an overall accuracy of 95%. Even after attaining such promising result, this model was unable to map even a single species across the study area from the many other unclassified tree species present in the study area (Clark et al., 2005). To overcome these problems, another approach has been adopted, namely to classify only single tree species. This method focuses only on one species which requires less training data (Liu et al., 2003). Therefore, single species classification has been performed for mapping tree species in low diverse ecosystems (Feret & Asner, 2012; M. Foody et al., 2005). Similar to the single tree species mapping approach, Baldeck & Asner, (2015) have mapped three species in the tropical forest of Barro Colorado Island, Panama which has a dense canopy. They made use of hyperspectral data obtained by High-fidelity Imaging Spectrometer (HiFIS) sensor having a spatial resolution of 1.12 m and spectral resolution of 9.4 nm using various support vector techniques.

While mapping of individual tree species, background signal such as undergrowth vegetation and bare soil cause variations in spectral signatures of the same species in a different location. This happens usually in

case of tropical forests. To overcome this problem reference data has been derived from the dense canopy of single species (Carleer & Wolff, 2004; Ghosh, Fassnacht et al., 2014).

2.3. Methods used for tree species classification

Classification of hyperspectral imagery has been attempted with traditional multispectral classification algorithms, and further modifications have also been done to improve the performance of the classifier for the hyperspectral data in terms of accuracy and robustness. Chutia et al. (2016) provided a detailed study of different classification techniques for hyperspectral data. Maximum Likelihood Classifier (MLC) based classification has been used in the past to classify hyperspectral imagery for tree species mapping after applying PCA to reduce the dimensionality of hyperspectral data that is to reduce the number of bands by removing the bands with high correlation. This was done to reduce the redundancy in the data (Jia & Richards, 1994). But MLC is biased for small training samples and cannot be applied for high dimensional data. To reduce the dimensionality of the data, some other techniques can also be implemented like the removal of individual bands, vegetation indices calculation, PCA or similar other transformations to the data. Some studies made use of Minimum Noise Fraction (MNF) transformation for the purpose of diversity mapping using imaging spectroscopy (Ghosh et al., 2014; Laurin et al., 2014; Leutner et al., 2012). But in MNF, PCA is applied for the two times due to which there are chances to lose even the important information (Richards & Jia, 2006).

Studies have been done where mapping of tree species has been carried out using SVM classifier; These studies provided good results in a closed-canopy and diverse tropical forest using hyperspectral data (Baldeck et al., 2015). The advantage of SVM is that it performs well even with the small training samples. But the selection of the kernel parameter is the major limitation of this approach.

For tree species identification, ensemble classifiers showed good performance (Ferreira et al., 2016). Ensemble classifiers are gaining popularity in recent time. There are three main approaches to construct the classifier ensemble that are bagging (Breiman, 1996), RF (Breiman, 2001) and boosting (Freund & Schapire, 1997). Bagging gives high accuracy but leads to low diversity in the individual classifier. RF is a version of bagging to enforce diversity among base classifier which improves accuracy of the classifier. Boosting is used to enhance the performance of the classifier. For hyperspectral data, RF has been used in many studies due to its ability to deal with an increasing number of input features. But in such cases where a large number of unimportant features are present, sometimes the performance of RF reduces. So the modification of RF gained popularity to deal with the high dimensional dataset. In a study done by Rodriguez et al. (2006), ensemble classifier was constructed by the modification of RF which is known as RoRF. The modification was based on a feature extraction technique that is PCA. In this method, each decision tree is trained in new different rotated spaces which gives increased individual classifier accuracy and diversity simultaneously.

Application of PCA is not suitable for feature extraction in the whole dataset (Heijden et al., 2004; Webb, 2002) because it results in loss of some relevant information. This is due to the reason that when PCA is performed in the whole data set and only a few components are retained then there are chances that more important components which correspond to small variance will be discarded. But PCA has performed better when applied for transformation of the data at each node in ensemble classifier. In a study conducted by Skurichina & Duin (2005), they proposed an ensemble classifier which was built by using

the PCA in the whole dataset. They have proved that a PCA based ensemble gave better results than ensembles based on random feature selection.

Tumer & Oza (2003) have used PCA as a dimensionality reduction tool for the generation of the ensemble. The number of classifier in their ensemble was equal to the number of classes in their study. Different sets of extracted features were selected to improve the diversity among the classifiers. Then for the training of each base classifier, PCA is applied on data of each class. The number of principal components retained was the parameter of the algorithm. This transformation was applied on the whole dataset, and each classifier was trained on the selected extracted features which were further used to distinguish the original classes. This is the reason to choose the size of ensemble the same as the number of classes present.

To analyse the performance of PCA based ensemble classifier, Rodriguez et al. (2006) performed a study in which they applied PCA as feature extraction technique to the feature subsets and reconstructed a feature set for each classifier in the ensemble. They considered all the principal components to avoid the loss of information. They called this a rotation forest and compared it with Bagging, Adaboost and RF. For all the methods they used the same number of classifiers in the ensemble. They performed this experiment on 33 datasets from the UCI Machine Learning Repository and showed that rotation forest gave the best results among all other methods. The reason for its good performance was the increased diversity among the classifier which was due to different feature subsets. In this study, they only considered PCA for the feature extraction.

Zhang & Suganthan (2014) have compared different method where different feature extraction techniques were used to transform the data at each node. They used PCA and LDA to see the performance and compared it with standard RF. In this experiment, the parameter which controls the size of the feature subset was fixed as default that is the square root of the total number of features in the dataset for all three classifiers so that a comparative study among the classifier can be possible. Also, they considered the same number of base classifiers which was 100 for all the method so that a fair comparison could be possible. Both PCA based RoRF and LDA based RoRF performed better than standard RF. It was because when data was transformed at each node, it increased the diversity among the decision trees and hence results in low correlation among them. In this study PCA based RoRF outperformed LDA based RoRF. It is because all the principal components are retained to preserve the variability information of the data, and the whole dataset was used to train each base classifier. Hence accuracy has increased.

.

2.4. Comparitive performance of classification algorithms

Classifier comparison is an important step to know whether the classification results are significantly different from each other. For this, the selection of suitable statistical test is required. In those cases where only a single test dataset can be used to evaluate the classification algorithms, and it is not possible to apply the test in which evaluation is done repetitively by making use of resampling technique like k-fold cross-validation. For such cases, Dietterich (1998) has suggested a McNemar test because it gives the low type I error. The type I error can be defined as the probability of wrongly detecting a difference among classifiers where no difference is present.

McNemar test has been used to compare the performance of 5 classification algorithms that are Bayes Net, IBK, Naive Bayes, J48 and Multilayer Perceptron (Bostanci & Bostanci, 2013). In this study, they experimented to justify the integrity of the McNemar test. For that, they compared the McNemar test with Kappa statistic and Root Mean Squared Error (RMSE) and found that McNemar test conforms to the Kappa statistic and RMSE. Cortés Rodríguez (2014) performed Land Use Land Cover (LULC) classification by using seven different ensemble classifiers and compared these classifiers by using the McNemar test.

3. STUDY AREA AND DATASETS

3.1. Study area

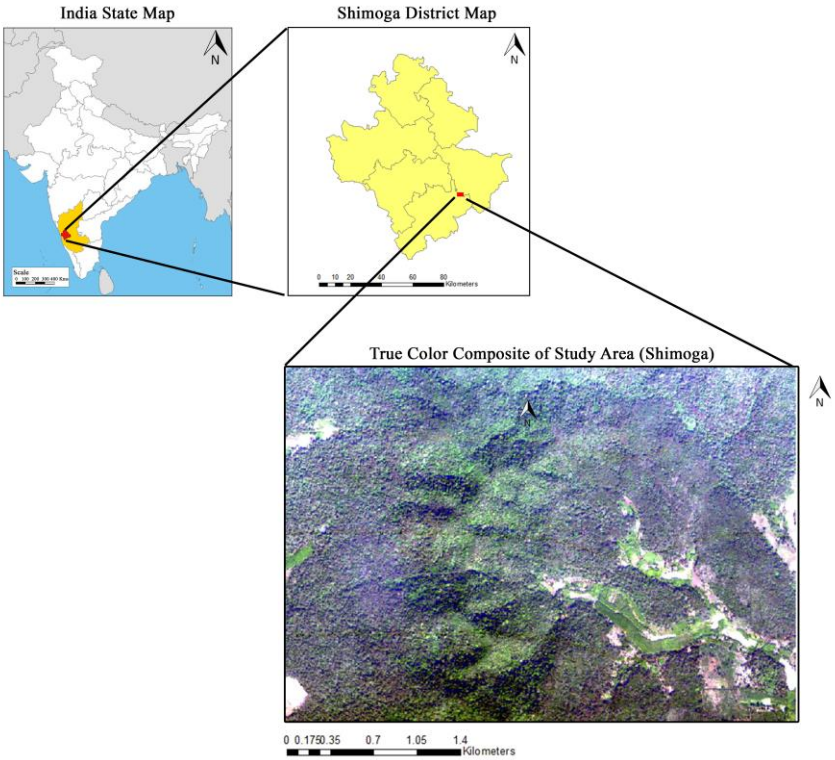


Figure 3.1 Study area. True colour composite (TCC) of AVIRIS-NG image (R: 54; G: 36; B: 18) over the Shimoga forest area (Karnataka)

The peninsular region of India, particularly the Western Ghats, is home to one of the eight hottest hotspots having rich biological diversity in the world and is declared as World Heritage Site by United Nations Educational, Scientific and Cultural Organisation (UNESCO) (The Times of India, 2012). The forests of Western Ghats lie within 12°N to 14°N covering areas of Coorg district, Hassan, Chikmagalur, Shimoga up to the southern region of Uttara Kannada. This research is targeted in Shimoga region which is the gateway to the hilly region of the Western Ghats Figure 3.1. The extent of the study area lies within 75.408949 to 75.446867 in longitude while 13.819849 to 13.846069 in latitude. The region has rich species diversity which is due to the tropical climate and heavy precipitation (Bhat et al., 2012). Rainfall season is from June to September, with maximum rainfall reported in July. And the reported driest month is March. Following Figure 3.2 shows the False colour composite (FCC) of vegetation of the study area during the driest month, i.e. March and the month just after the rainfall season gets over i.e. November. This FCC was analysed by using Sentinel data to see the clear spread of the vegetation. During dry season, evergreen

species do not shed all of their leaves like deciduous species. So during the dry month also, some part in the Figure 3.2 is showing red colour in FCC which denotes that this region has the dominance of evergreen species. A transition from moist deciduous to pure evergreen can be clearly seen in the study area.

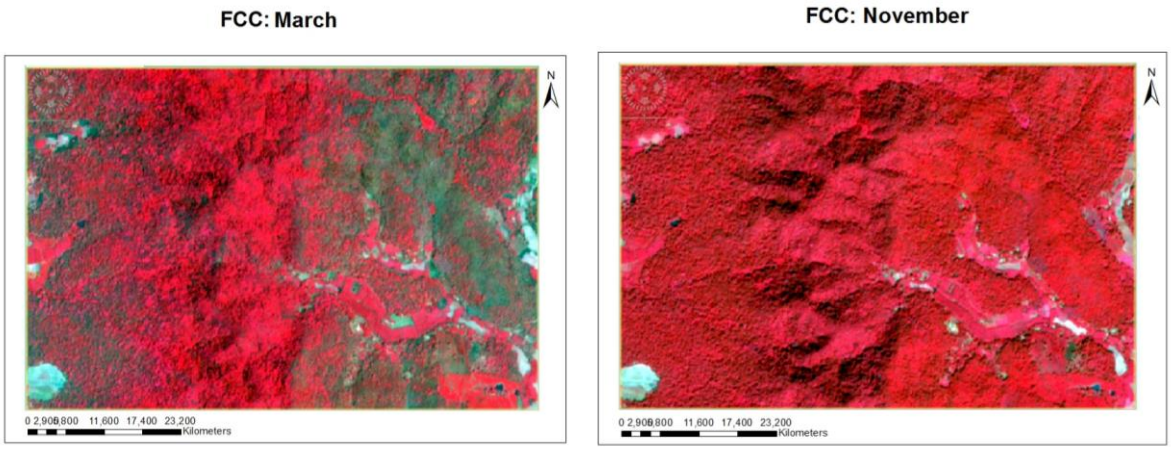


Figure 3.2: Analysis of vegetation in Shimoga, Karnataka by using False Colour Composite (FCC) images of Sentinel 2 data

According to Champion & Seth (1968), this area consists of Southern Tropical Semi-Evergreen Forest (2A/C2) and South Tropical Moist Deciduous Forest (3B/2). Eastern side of the study area has degraded forest, so no proper shape and form is present. Even plantation species like Teak & Eucalyptus are present inside the natural forest of this region. Evergreen tree species are mostly present in the western side of the study area, primarily concentrated in the mountain range named Mandgadde. Deciduous species are spread throughout the study area because, in the evergreen region, past illegal felling and fire have left open patches in which restoration has been done with deciduous species as well.

3.2. Dataset used

Airborne Visible/Infrared Imaging Spectrometer-Next Generation (AVIRIS-NG), developed by Jet Propulsion Laboratory (JPL) of The National Aeronautics and Space Administration (NASA) is used in this study. It is an airborne hyperspectral sensor having wavelength range 376-2500 nm with a narrow bandwidth of 10 nm, 425 bands and 5m spatial resolution (gisresources, 2013). The swath width of AVIRIS is around 11 km. The data acquisition was made on 1st January 2016. The dataset was available in the Visualisation of Earth observation Data and Archival System (VEDAS) of Space Application Centre (SAC) (Vedas SAC, 2016) as shown in Figure 3.3. For this study, 12 sqkm area has been selected for analysis.

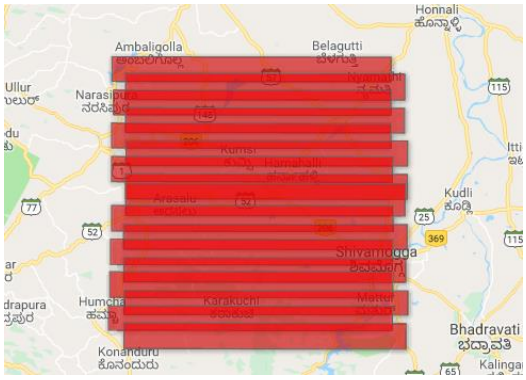


Figure 3.3: AVIRIS-NG strips available for Shimoga, Karnataka (Vedas SAC, 2016)

Table 3.1: Dataset information

Dataset information	
Name	AVIRIS-NG
Sensor	Hyperspectral
Spatial resolution	5m
Spectral resolution	425
Swath	11km
Wavelength range	376-2500 nm
Bandwidth	10nm
Source	(Vedas SAC, 2016)

3.3. Software used

Following software have been used in this study:

1. All the classification methods and their comparison were done using Python 2. The libraries used were: gdal, ogr, numpy, pandas, researchpy, sklearn (Pedregosa et al., 2011).
2. R-Studio (R Development Core Team, 2010) was used for field plot generation by using clhs package and for calculating covariance between different species by using corrplot package (Wei & Simko, 2017) .
3. ENVI Classic 5.0 was used for visualisation and pre-processing of data and spectral signature generation.
4. ArcGIS (ArcMap v 10.1) developed by ESRI was used for data preparation and map generation.
5. QGIS Desktop version 3.4.3-Madeira (QGIS Development Team, 2009) was used for digitization of tree crown and tree species measurement.
6. All the three algorithms were computational intensive; hence High-performance computing (HPC) systems were utilised for their processing.

4. METHODOLOGY

This chapter describes the methods adopted to achieve the objectives of this research. The methods include the hyperspectral data preprocessing, sampling to generate field plots, collection of field data and its analysis, classification methods SVM, RF and RoRF for tree species mapping, accuracy assessment and the classifier comparison, tree species richness estimation both from the field and from the classified imagery. The general methodology has been highlighted in Figure 4.1.

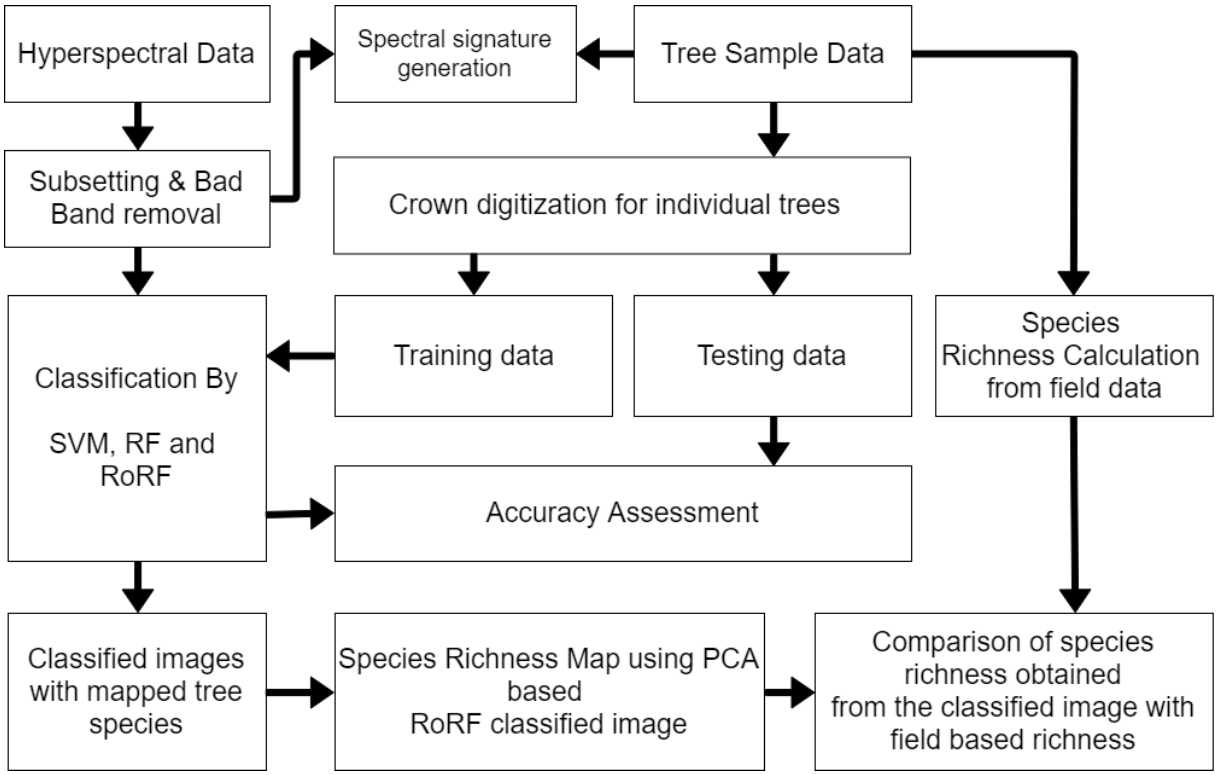


Figure 4.1: Generalized Methodology

4.1. Data preprocessing

The available data of AVIRIS-NG was of Level-2 (L2) (reflectance data) which means it was already radiometrically and atmospherically corrected. L2 data is produced at the Jet Propulsion Laboratory (JPL) Science Data System (SDS). All the bands are visualized individually, and water absorption bands and noisy bands are removed from the dataset by using the tool Resize data (Spatial/Spectral) in ENVI Classic image processing software version 5.0. Table 4.1 shows the list of the bad bands that were removed. The area outside the natural forest is cropped from the study area by using tool Subset Data via (region of interest) ROIs in ENVI. A total of 367 bands were retained which were further used for classification

Table 4.1: List of bad bands

AVIRIS-NG Bands	Band Wavelength (nm)	Remarks
1-10	376.44 to 421.52	Noisy bands
195-207	1348.12 to 1408.23	Water vapour absorption bands
287-316	1808.92 to 1954.17	Water vapour absorption bands
325-329	1999.25 to 2019.28	Noisy bands

4.2. Field data collection

The data collected from the field should be capable of representing the whole study area. Different criteria were considered in several studies for the collection of tree sample data. Clark & Roberts (2012) collected reference data according to visual interpretation like dominant species, isolated species etc. Jensen et al. (2012) considered accessibility in the study area. Some studies considered a group of species or homogenous crown (van Aardt & Wynne, 2007) and reducing background signal by selecting only the dense crown (Youngentob et al., 2011). Some studies have attempted to consider understory trees which have proved to be a difficult and challenging task (Korpela, Hovi, & Morsdorf, 2012). So it has been suggested for small trees, to do area-based classification rather than single species based approach. Only some studies considered the increasing representation of each species (Engler et al., 2013). In a study done by Leckie et al. (2005), several reference classes have been taken, which later merged into a single class to increase representation in each class and has proved to give good results.

The airborne hyperspectral data was collected on January 1, 2016, but field data collection was done in November 2018. It is assumed that there will be no significant change in tree species distribution between image acqusition and ground truth collection. Field data collection was done for 12 days from November 9, 2018, to November 20, 2018.

A GPS-Aided GEO Augmented Navigation (GAGAN-GPS), named "Parishudha" was used to collect the location data of the tree species. It has an accuracy of 0.5m to 2m. Inch tape and ropes were used for laying plots. The magnetic compass was used to lay plots with proper orientation.

Figure 4.2: GPS-Aided GEO Augmented Navigation (GAGAN-GPS) "Parishudha"

A Latin square is a square grid which contains only one sample in each row and column. When a Latin square is generalised to a "n" number of dimensions and in each axis-aligned hyperplane there is only one sample present is called the Latin hypercube (Figure 4.3) (Minasny & McBratney, 2006). In this study, Conditioned Latin hypercube sampling (cLHS) was used to generate field plots. The cLHS is a type stratified random sampling method. In this, sampling is done using some ancillary information. In this study, it was decided to generate 30 sample plots. Two variables that are Topographic Wetness Index (TWI) and Aspect has been used for ancillary information. These parameters were selected because species distribution mainly depends on soil quality, aspect, topography and drainage in the area (Moeslund et al., 2013). Both TWI and the aspect were calculated by using the Cartosat Digital Elevation Model (DEM) of 30 m spatial resolution.

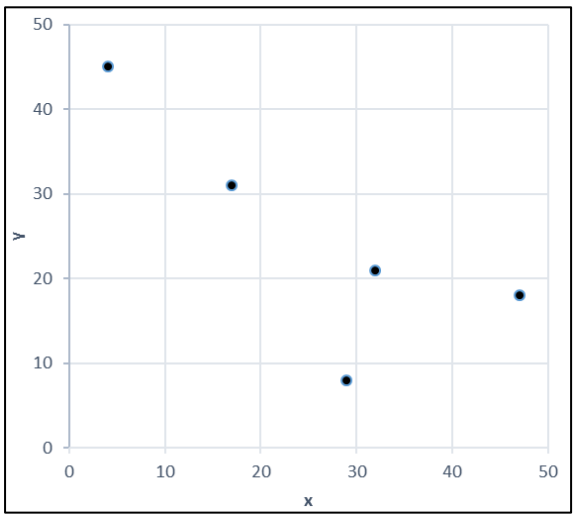


Figure 4.3: An example of 2-D Latin hypercube. Here, x and y are the two variables divided into equal classes of equal interval. The dots are the sample points. Note that for each class in both the axis, one sample point is present.

TWI (Beven & Kirkby, 1979) is a function of slope and flow accumulation of an area. It has shown a good correlation with species richness (Song & Cao, 2017). TWI was calculated in Arc-GIS 10.1, using python based tool provided by Fricker, (2017). It is given by (Equation 4.1):

$$w = ln\left(\frac{\alpha}{tan\beta}\right) \quad \text{(Equation 4.1)}$$

Where,

w = topographic wetness index

α = local area of flow accumulation

β = local slope (in degrees)

In the cLHS method, the range of selected variable (TWI and aspect) was divided into 30 equal probable intervals. And then in each Latin hypercube, one sample point was laid. As a result, 30 sample points were obtained. Note that the number of division is equal to the number of the sample points required. The advantage of this technique is that the number of samples does not depend upon the variables used (Jenkins, 2015). Because from the Figure 4.3 it can be seen that the two variables (x and y) are divided into 5 classes for generating 5 sample points. Similarly in this study, the two variables (TWI and aspect) were classified into 30 parts to generate 30 sample points. Conditioned Latin Hypercube Sampling (CLHS) library (Roudier, 2011) has been used in RStudio to perform cLHS sampling, provided with the number of required sample points, i.e. 30. Figure 4.5 shows the sample plots from where the tree species location data has been collected. A total of 30 sample plots were decided but on field it was found out that the 5 points were inaccessible due to the rough terrain and security issues. So, those points were rejected and tree location data from 25 plots was collected.

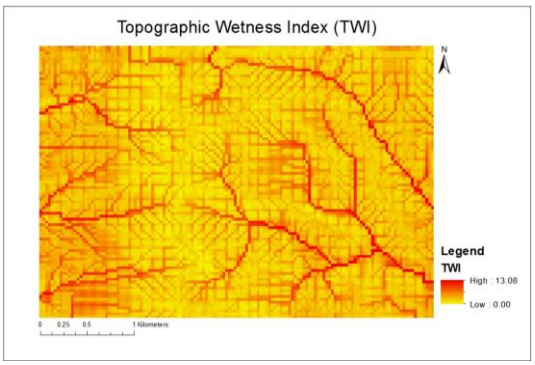


Figure 4.4: Topographic Wetness Index (TWI)

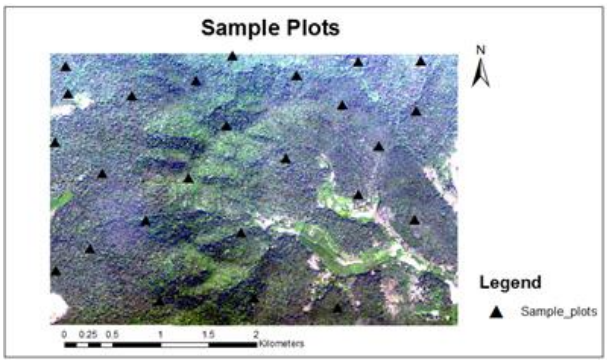


Figure 4.5: Sample plots

For each plot, location of all the four corners of plots was measured using the "Parishudha" with accuracy ranging from 0.5 m to 2 m. Since ranging rods were not available, so a right-angled triangle (Pythagoras theorem) was made at the corner of the plot for proper orientation of plot with the help of inch tape and a magnetic compass Figure 4.6 .Location of all trees within the plot was noted. Group of same tree species and dense crowns were given special importance so that for generation of spectral signatures background noise and mixed crowns could be avoided. In total 320 tree location points were collected representing 20 tree species.

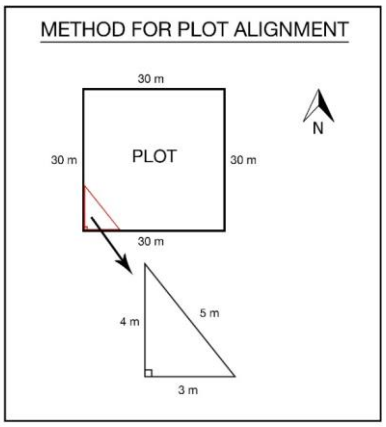


Figure 4.6: Field plot alignment

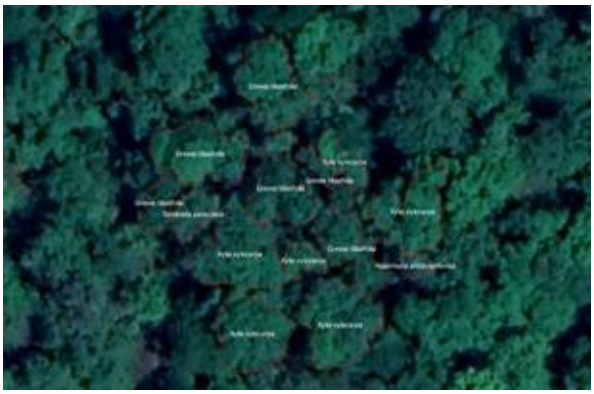


Figure 4.7: Digitization of tree crowns over world view image (Google Earth, 2019)

4.3. Data preparation

The collected 320 points shapefile was imported in Quantum GIS (QGIS) software to digitise the crowns of trees. Google base map was used to delineate the tree crown using the QGIS plugin QuickMapServices. As a result, a total of 320 polygons consisting of 1,738 pixels in total were obtained (Figure 4.7). These polygons were divided into a training set and testing set manually, considering that training and testing polygons of each class should be selected from different plots. Following is the Table 4.2 showing the training and testing data.

Table 4.2: Tree species identified in the study area

Species	Code	Training data		Testing data	
		Polygon	Pixels	Polygon	Pixels
Alsodephne semicarpifolia	AS	7	22	4	41
Bahunia racemosa	BR	8	33	3	29
Bombax ceiba	BC	4	17	3	26
Cassia fistula	CF	7	31	3	33
Dalbergia latifolia	DL	8	41	3	44
Delinia pentagyna	DP	7	49	3	20
Eucalyptus grandis	EG	9	16	4	14
Grewia tiliaefolia	GT	11	63	6	53
Holarrhena antidysenterica	HA	14	61	7	24
Kydia calycina	KC	6	23	3	21
Lagestormia lanciolata	LL	12	49	6	94
Lagestormia parviflora	LP	19	33	4	30
Lannea grandis	LG	17	104	7	37
Melia dubia	MD	4	28	3	26
Phyllanthus emblica	PE	8	63	3	20
Tectona grandis	TG	8	20	3	41
Terminalia bellarica	TB	8	41	4	53
Terminalia paniculata	TP	30	75	10	95
Terminalia tomentosa	TT	8	57	4	39
Xylia xylocarpa	XX	34	104	8	68
	Total	**229**	**930**	**91**	**808**

4.4. Classification of hyperspectral imagery

4.4.1. Support Vector Machine (SVM)

The SVM was developed by Vapnik (1995). This classifier tries to find the optimal hyperplane (a decision boundary) in multidimensional space in such a way that there will be a maximum margin between the classes to minimize the structural risk (classification errors). The hyperplane is built on the basis of training data properties which tends to maximise the margin of separation between the different classes. A subset of training data contributes in constructing of the hyperplane and called as support vectors. Classes can be clearly separable or non-separable. As in remote sensing classes are generally not able to separate linearly; non-linear SVM can be implemented. In non-linear SVM, raw input data or feature vector ($x_i \in R^d$) is mapped into a higher dimension space $\mathcal{H}$ to improve the class separability. It facilitates the fitting of the linear hyperplane. The training samples are projected in higher dimension $\mathcal{H}$ by means of a nonlinear vector mapping function called Φ: $R^d \rightarrow \mathcal{H}$ by a decision rule shown in the (Equation 4.2.

$$f(x) = sign\left(\sum_{i=1}^{nsv} \alpha_i \, y_i \Phi(x).\Phi(x_i) + b\right) \qquad \text{(Equation 4.2)}$$

Where, $\{x_{i,}y_i\}$ is the training dataset in which x_i is the observed features while y_i is the label information of the class. α_i is the positive Lagrange multiplier for each training point ($\alpha_i > 0$ are support vectors). nsv is the number of support vectors, x is the point lying on hyperplane, b is the bias which is computed for stability concerns by using all the support vectors on the margin.

In high dimension space $\Phi(x) * \Phi(x_i)$ can be computationally intensive so Vapnik, (1995) proposed a kernel function denoted by $K(x, y)$. This function used in training algorithm and reduces the computational burden. The generalised decision rule is given in (Equation 4.3)

$$f(x) = sign\left(\sum_{i=1}^{nsv} \alpha_i \, y_i K(x, x_i) + b\right) \qquad \text{(Equation 4.3)}$$

Some of the common kernel functions are linear, polynomial (homogeneous), polynomial (inhomogeneous), Radial Basis Function (RBF), Gaussian Radial Basis Function and sigmoid. In this study RBF kernel function is adopted which is given by the (Equation 4.4):

$$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2), \gamma > 0 \qquad \text{(Equation 4.4)}$$

In RBF two parameters have to be optimised before performing the training stage. These are the penalty parameter denoted as C (error term) and kernel parameter denoted by γ (gamma). The C parameter has to be optimised in case of all the SVMs because it controls the trade-off between decision rule complexity and training error frequency. While γ needs to be defined before applying RBF-SVM. In this study, the RBF kernel is selected because the linear kernel was not able to handle nonlinearly separable classes. The polynomial and sigmoid kernels needs more parameter to tune than the RBF kernel also the computation is more stable in RBF kernel (Tso & Mather, 2009).

RBF-SVM classification is performed using 'svm' function present in sklearn library for Python 2. SVM is trained using the optimal value for C and gamma. The optimal C value and gamma value was chosen by applying loop for the C value range (1 to 1000 at an interval of 100) and gamma value range (0.1 to 1 at an interval of 0.1). Maximum accuracy was obtained at C= 100 and gamma=0.6. Further accuracy assessment was done using the confusion matrix.

4.4.2. Random Forest (RF)

The RF (Breiman, 2001) is a machine learning algorithm which consists of a combination of decision trees classifiers. In this, each classifier is generated by selecting a set of independent random samples from the training set of the input vector and forming a forest. Bagging method is used to generate forest randomly. Bagging avoids overfitting and also improves the classification accuracy (Breiman, 1996). Bagging generates a random sample with replacement of size n from the training set N and makes a new training set (where,$n < N$). Suppose data contains M attributes (spectral bands), so m (where,$m < M$) attributes are also randomly selected for each node to provide the base for the best split at that node of the tree.

So there may be the chance that some samples are selected many times while some may not choose at all. Approximately two-third samples used for training called as in-bag samples while remaining are called as out-of-the bag samples, which are used in internal cross-validation to estimate the performance of resulting RF model. This error estimate called as out-of-bag (OOB) error. The number of decision trees (Ntree) is defined by users, and each decision tree is produced independently without pruning. Each node of the tree splits using the number of features (Mtry) that is a user-defined parameter. The algorithm then creates decision trees with high variance and low bias. The final classification decision is taken considering arithmetic mean of the class assignment probabilities calculated by all the trees in the RF. Then the unlabelled data is given as input and is evaluated against all the decision trees in classifier ensemble. Finally voting is done for class membership by each tree and the class with maximum votes is finally assigned (Belgiu & Drăguţ, 2016).

RF is performed using 'RandomForestClassifier' function present in the ensemble class of sklearn library for Python 2. Size of the feature subset was fixed at 19 (square root of the number of features), and Ntree was defined as 500 since the errors get stabilize at this number of decision trees (Belgiu & Drăguţ, 2016).

4.4.3. Rotation Random Forest

The RoRF (Rodriguez et al., 2006) is a classifier ensemble in which the original data feature space is transformed into another feature space. This method is rather different from the traditional methods, where, whole data is transformed using feature reduction techniques before the classification. Instead, in RoRF, data transformation applies at each node in different subspace. In this method, the feature set is split into K subsets, and in each subset data transformation algorithm is applied, and then the new extracted feature set is rearranged while keeping all the components. It results in increasing both member diversities and individual accuracy within the classifier. In this study, PCA based RoRF has been used where PCA is used to transform data at each node.

The theory behind the PCA based RoRF:
Assumptions:-

1) $x = [x_1, x_2, \dots, x_n]^T$ dataset having n features
2) X = Training dataset in the form of $N \times n$ matrix
3) $y = [y_1, y_2, \dots, y_n]^T$ It is the vector with class labels in the form of $N \times 1$ matrix and Class labels are denoted by the set $\{w_1, w_2, \dots, w_c\}$, where c is the total no. of class
4) $L = D_1, D_2, \dots, D_L$ are the classifier ensemble where L is the total no. of decision trees
5) F is the feature set

To formulate Base classifiers:-

- All classifiers are trained in parallel for $i = 1, \dots, L$
- Split feature set (F) into F_{ij} disjoint subsets: for $j = 1, \dots, K$, where K is the total number of subsets. Each subset contains M features.
- Now, consider a random non-empty subset, X_{ij} of classes (for given dataset X) for the feature set $F_{i,j}$ at $j = 1, \dots, K$. (Note: the Eigen values of the considered subset may contain zero values)
- Reject the terms which are having zero Eigenvalues in X_{ij} (so that, $M_j \leq M$)
- For each subset$X_{i,j}$, a bootstrap sample is selected i.e. 75 % of the data count (X_{ij}) is selected then the new set of $X_{i,j}$ is denoted as $X'_{i,j}$
- Apply PCA on $X'_{i,j}$ and F_{ij} and matrix of coefficient [C_{ij}] is obtained, For $j = 1, \dots, K$
- Arrange coefficient C_{ij} matrix in Rotation matrix (R_i) format (Equation 4.5) and rearrange the column(R_i) into R_i^a (size N x n) format to match the order of feature set (F).

$$R_i = \begin{bmatrix} a_{i,1}^{(1)} a_{i,1}^{(2)}, \dots, a_{i,1}^{(M_1)} & 0 & \cdots & 0 \\ 0 & a_{i,2}^{(1)} a_{i,2}^{(2)}, \dots, a_{i,2}^{(M_2)} & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & a_{i,K}^{(1)} a_{i,K}^{(2)}, \dots, a_{i,K}^{(M_K)} \end{bmatrix} \quad \text{(Equation 4.5)}$$

The training classifier, D_i is constructed using the training set = $(XR_i{}^a, Y)$

Classification phase:-

For a given unlabelled input data x, let $d_{i,j}(xR_i{}^a)$ be the probability that classifier D_i has assigned by stating the hypothesis that x belongs to class w_j

Now Confidence of each class w_j can be calculated by average combination method (Equation 4.6).

$$\mu_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(xR_i{}^a) \text{ where } j = 1, \dots, c \qquad \text{(Equation 4.6)}$$

x will be assigned to the class which will have maximum confidence.

In this study, RoRF algorithm was applied to the AVIRIS NG hyperspectral imagery. RoRF was performed using sklearn library for Python 2. Two key parameters of RoRF that are the number of features in each subset (M) and the number of decision trees (L) and since in this case number of features were 367, so M was fixed to 19 (square root of the number of features). Decision trees were selected as the base classifier and the number of trees (L) were selected as 500. These two values were taken same for both RF and RoRF, so that a fair comparison could be possible. Overall accuracy was estimated using the confusion matrix to assess the performance of the classifier.

4.4.4. Accuracy assessment of the classification algorithms

Accuracy assessment for tree species richness is an important step to assess the performance quantitatively. This study has utilised a confusion matrix to determine the classification accuracy from the validation samples which has further helped to derive Kappa index, overall accuracy, producer accuracy and user accuracy (Congalton, 1991). Confusion matrix, which is also known as the error matrix is formed for comparison or cross-tabulation of the number of pixels of all classes in the classified image to the actual class (obtained from testing data). Table 4.3 shows the layout of the confusion matrix.

Table 4.3: Layout of a confusion matrix

Testing data	Classified image				Total
	Class 1	Class 2	……	Class n	
Class 1	a_{11}	a_{12}	……	a_{1n}	R_1
Class 2	a_{21}	a_{22}	……	a_{2n}	R_2
……	……	……	……	……	……
Class n	a_{n1}	a_{n2}	……	a_{nn}	R_n
Total	C_1	C_2		C_n	

Here, a_{pp} are the number of pixels that are correctly classified and a_{pq} are the number of pixels which belongs to class p in testing data but it is classified into class q.

Overall Accuracy is calculated by dividing the sum of the count of the correctly classified pixels (diagonal elements of the confusion matrix) by the total count of pixels in the testing data, as shown in (Equation 4.7)

$$OA = \frac{\sum_{i=1}^{n} a_{ii}}{\sum_{i=1}^{n} R_i} \quad \text{(Equation 4.7)}$$

Producer Accuracy is calculated by dividing the total number of correctly classified pixels in class i by the total number of pixels present in that class in testing data. Given in the (Equation 4.8). This is a measure of error of omission, it indicates the probability that a pixel is correctly classified this measure is also known as the error of omission.

$$PA_i = \frac{a_{ii}}{R_i} \quad \text{(Equation 4.8)}$$

User accuracy is calculated by dividing the total number of pixels that are correctly classified in the class i divided by total number of pixels that are classified in that class. Given in the (Equation 4.9) Since it indicates the probability that a pixel classified on the image is actually represent that class on the ground, this measure is also known as the error of commission.

$$PA_i = \frac{a_{ii}}{R_i} \quad \text{(Equation 4.9)}$$

In this study accuracy of the PCA-based RoRF, RF and SVM was assessed by using sklearn library using "classifier.score" function and kappa score was calculated by using "cohen_kappa_score" function in Python 2.

4.4.5. Comparison of classification algorithms

Classification algorithms (SVM, RF and RoRF) are compared by using the McNemar test. The null hypothesis for this test was formulated as both the classifiers are not significantly different from each other. Alternate hypothesis was that the performance of both the classifiers was different from each other. z Score has been calculated by using Equation 4.7. In case of the $z = 0$, it means both the algorithms have performed the same. At the significance level of $\alpha = 0.05$, p value is calculated by using Equation 4.8. If the p value is lower than the selected significance level, then reject the null hypothesis that both the classifiers are performing the same. Hence, both the classifiers are significantly different from each other vice versa (Bostanci & Bostanci, 2013).

McNemar test is based on the 2x2 contingency matrix of the algorithm predictions. Following table 4.4 shows the possible outcomes of the algorithms.

The test statistic for McNemar test is given by the (Equation 4.10):

$$z = \frac{(|N_{cw} - N_{wc}| - 1)}{\sqrt{N_{cw} + N_{wc}}} \quad \text{(Equation 4.10)}$$

Where,
N_{cw}: denotes the number of times the algorithm 1 classified correctly and the algorithm 2 failed.
N_{wc}: denotes the number of times the algorithm 2 classified correctly and the algorithm 1 failed
N_{ww}: denotes the number of times when both the algorithms failed
N_{cc}: denotes the number of times when both the algorithms classified correctly

The p value is given by the
(Equation 4.11):

$$p = 2 \sum_{i=N_{cw}}^{n} \binom{n}{i} 0.5^i (1 - 0.5)^{n-i} \quad \text{(Equation 4.11)}$$

Here, n denotes the total number of samples, $n = N_{cw} + N_{wc}$

Table 4.4: Layout of the contingency matrix for McNemar test

	Algorithm 1 (wrong)	**Algorithm 1 (correct)**
Algorithm 2 (wrong)	N_{ww}	N_{cw}
Algorithm 2 (correct)	N_{wc}	N_{cc}

In this study, McNemar test was calculated using "researchpy" library in Python 2. z Score was calculated between two classification algorithms at a time. Hence, this test was applied three times to compare the performance of all three combinations of classifiers.

4.4.6. Assessment of tree species richness

In principle, species richness is the total number of unique species in a particular area. Species richness was calculated by using TomBio tool (Burkmar, 2015) in QGIS. This tool can be used for obtaining the distribution and abundance of the living organisms. It was originally made to use in the United Kingdom but can be used in any geographical region. In this tool, the grid is created and the size of grid is user-defined, to obtain the information regarding biological records within each of the grid. This tool counts the total number of points and the unique points in each grid. In the output attribute table the total number of points is represented as "abundance" and the total count of unique pixels is represented by "richness", along with the spatial information of the grid.

The TomBio version 3.1.1 is the latest version of the tool and is available from the QGIS plugin repository. It takes .csv file as input. Therefore RoRF classified image was converted into point layer and saved in csv format. This file was imported in tool and latitude, longitude and taxon column were specified. Since the size of field plot was 30m x 30m, the grid size was also selected the same so that a proper comparison could be possible between species richness obtained from field and the RoRF classified imagery. This tool was run by selecting the *create map layer* button and the map layer was created having the information associated with each grid. That is the information about the number of unique pixels present in each grid and also the total number of pixels present in the grid. These can be visualised clearly in the attribute table. Here, a number of unique pixels are the different species in the grid. Hence, species richness per 30m x 30m was obtained. The richness map was then created using the Arc GIS.

The richness from the field was also calculated manually and then compared with the species richness obtained for the same region (field plot) from the RoRF classified map.

5. RESULTS

This chapter presents all the results and their analysis which were obtained by applying the methodology steps present in chapter 4 on the hyperspectral imagery. This chapter has been divided into following subsections:

- Spectral separability
- Classification results analysis
    - Result obtained from SVM
    - Result obtained from RF
    - Result obtained from RoRF
- Comparison of classifiers
- Forest type map
- Species richness

5.1. Spectral separability

Spectral signatures were generated from the pixel corresponding to the tree location point, preferably where tree species were present in groups. The spectral signatures of all the tree species are shown in Figure 5.1. This was done to see the spectral variability among the different tree species. For better clarity, covariance of spectral reflectance among the tree species was calculated, which is shown in Figure 5.1. The range of Covariance lies between 0 to 0.02, which is very low. Hence, the variability among the spectral signatures proved to be low. Among all the tree species classes, *Eucalyptus grandis* has showed maximum covariance with all other species. And *Lannea grandis* has the lowest covariance value with all other tree species. *Lannea grandis* shows little spectral variance with respect to *Eucalyptus grandis*, *Tectona grandis* and *Dalbergia latifolia.*

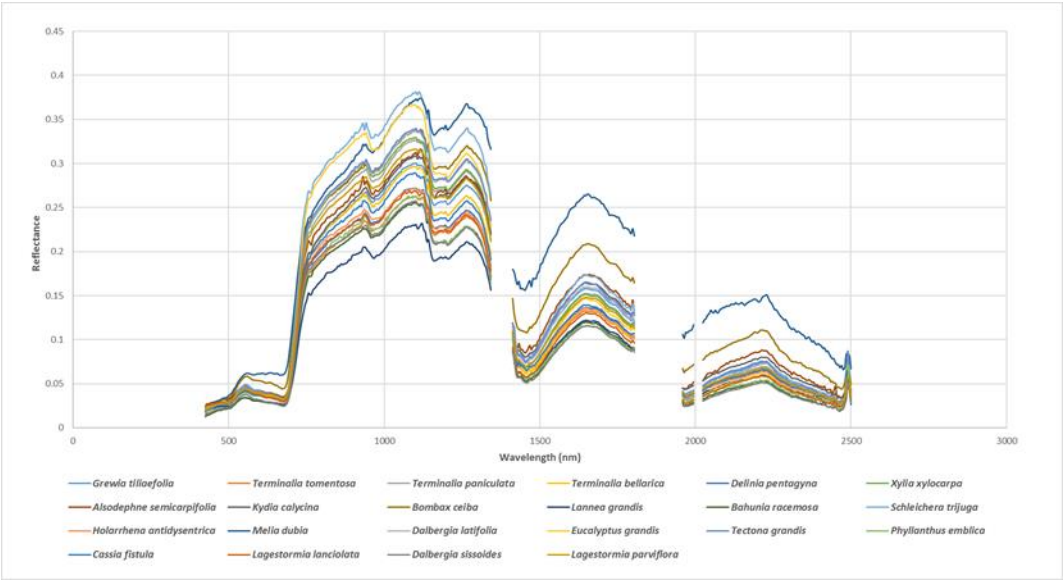
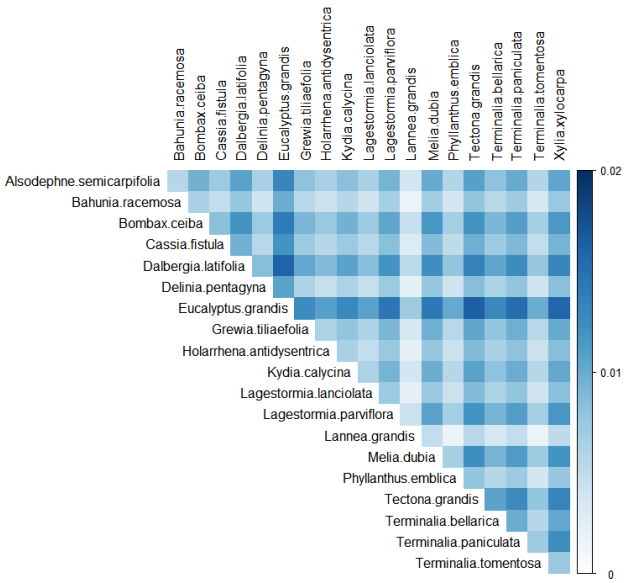


Figure 5.1: Comparison of Spectral signature between different tree species. The graph is plotted between wavelength and reflectance on x and y axis respectively. Different colours represent different tree species identified in study area.



.

Figure 5.2: Covariance between spectral signatures of the tree species. The covariance value increases from light colour to dark colour.

5.2. Classification results analysis

In this research three different classifiers, namely SVM, RF & RoRF have been used and compared so as to provide best tree species level classification based on AVIRIS-NG hyperspectral imagery for Shimoga region of Karnataka, India. The classifier with best performance shall be used to generate the final results for the tree species richness and analysis. This section explains the classification results obtained from each classifier.

5.2.1. Result obtained from SVM

The SVM algorithm was applied to classify the hyperspectral imagery of the Shimoga region in Karnataka. SVM consists of two major parameters, C value and Gamma, which needs to be optimised accordingly to obtain maximum classification accuracy. For applying parameter optimization, a range of values of C and gamma have been defined. The set of values which provided highest accuracy was used to perform classification. To determine these parameters, the range defined for C value was 1 to 1,000 at an interval of 100, while the range defined for gamma was 0.1 to 1 at an interval of 0.1. A nested for loop has been applied on the above mentioned ranges, which provided the set of C = 100 and gamma = 0.6 to give highest accuracy. The variation of accuracy with different set of C and gamma values has been shown in Figure 5.3.

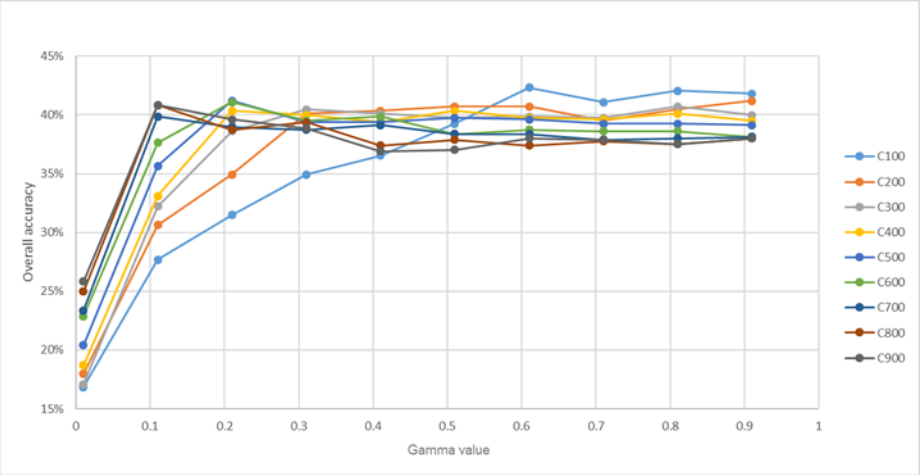
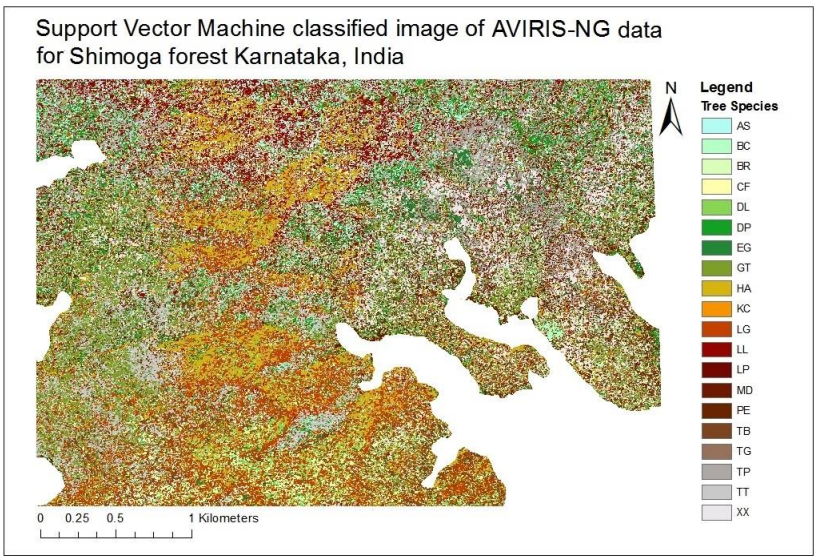


Figure 5.3: Variation in overall accuracy with gamma value at constant value of C

The classified image obtained using SVM with optimized parameters has been shown in Figure 5.4. From the classified image, it can be observed that *Bahunia racemosa*, *Dalbergia latifolia* and *Lannea grandis* are spread over southern part of the area. This region mostly consists of hilly area. It is a part of Mandgadde range. This range extends from north to south. On the peaks of this range *Lannea grandis* and *Holarrhena antidysenterica* were dominant species. In the flat region of southern part *Terminalia bellarica* has been seen more. In northern part more number of tree species classes were together showing more species richness in northern part. *Alsodephne semicarpifolia* and *Lagestormia Lanciolata* were in dominance while *Tectona grandis* was spread more in northern part but not in dominance, since *Tectona grandis* is not a native species of this area. It was planted to fill the blank patches inside the forest. In the western part of study area *Terminalia tomentosa* and *Grewia tiliaefolia* were together dominating the forest. *Delinia pentagyna*, *Grewia tiliaefolia* and *Phyllanthus emblica* were in flat portions of the study area. *Terminalia paniculata*, *Xylia xylocarpa* and *Lagestormia parviflora* are mostly present in the eastern side of the study area. Also, *Eucalyptus grandis* is seen in some patches together with them in eastern part of study area.



AS - Alsodephne semicarpifolia, BR - Bahunia racemosa, BC - Bombax ceiba, CF - Cassia fistula, DF - Dalbergia latifolia, DP - Delinia pentagyna, EG - Eucalyptus grandis, GT - Grewia tiliaefolia, HA - Holarrhena antidysenterica, KC - Kydia calycina, LL - Lagestormia lanciolata, LP - Lagestormia parviflora, LG - Lannea grandis, MD - Melia dubia PE - Phyllanthus emblica, TE - Tectona grandis, TB - Terminalia bellarica, TP - Terminalia paniculata, TT - Terminalia tomentosa, XX - Xylia xylocarpa

Figure 5.4: Classified output generated using Support Vector Machine

The confusion matrix obtained from the classified result has been shown in Table 5.1. The SVM gave an overall accuracy of 41.21% and kappa value of 0.37. From the confusion matrix, it can been observed that *Bahunia racemosa* has maximum producer accuracy i.e. 82.76% followed by *Eucalyptus grandis* and *Bombax ceiba* which has a producer accuracy of 78.57% and 73.08% respectively. For class *Lagestormia parviflora*, SVM failed to correctly classify due to overlapping of signatures with other species. Training data of *Lagestormia parviflora* has 33 number of pixels and mostly it is classified into two dominant species that are *Terminalia paniculata* and *Xylia xylocarpa*.

For *Cassia fistula* and *Xylia xylocarpa*, producer accuracy is very low i.e. 9% and 16.18% respectively. Rest of the species classes achieved producer accuracy between 20% to 70%. The dominant species *Xylia xylocarpa* is mostly misclassified with *Grewia tiliaefolia* and *Alsodephne semicarpifolia*. In case of *Lannea grandis* 17 out of 37 pixels were correctly classified by SVM. But 55 pixels from other species were misclassified with *Lannea grandis*, as a result 23.61% of user accuracy.

Table 5.1: Confusion Matrix for SVM

Ground refrence point \ Classified AVIRIS-NG image	AS	BR	BC	CF	DL	DP	EG	GT	HA	KC	LL	LP	LG	MD	PE	TG	TB	TP	TT	XX	Other	Total	Producer Accuracy(%)
AS	19	0	0	0	0	2	0	0	2	0	4	1	1	0	5	0	0	6	0	1	0	41	46.3415
BR	0	24	0	0	1	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	29	82.7586
BC	3	0	19	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	26	73.0769
CF	2	0	0	3	0	0	0	2	3	0	2	2	3	0	3	0	1	6	0	6	0	33	9.0909
DL	0	1	0	0	22	0	9	0	7	0	0	0	3	0	0	0	0	0	0	0	2	44	50.0000
DP	1	0	0	2	1	6	0	2	0	0	3	0	0	1	1	0	0	1	2	0	0	20	30.0000
EG	0	0	0	0	0	0	11	0	0	0	0	0	0	0	1	0	0	0	0	2	0	14	78.5714
GT	7	2	0	0	3	2	0	16	0	4	3	0	2	1	2	0	0	3	7	1	0	53	30.1887
HA	2	0	0	0	0	2	0	0	12	0	3	0	3	0	0	0	0	0	0	2	0	24	50.0000
KC	0	2	0	0	0	2	0	3	0	5	0	0	2	4	1	0	0	0	2	0	0	21	23.8095
LL	0	0	0	13	0	0	0	0	5	0	58	0	4	0	1	7	0	0	5	1	0	94	61.7021
LP	0	0	0	2	4	0	0	1	0	0	0	0	1	0	4	0	12	2	1	3	0	30	0.0000
LG	1	0	0	1	0	1	0	7	2	0	0	1	17	1	0	0	1	1	3	1	0	37	45.9459
MD	0	0	0	3	0	1	0	2	0	0	0	0	0	17	1	0	1	1	0	0	0	26	65.3846
PE	2	0	0	0	0	1	0	5	0	0	2	0	1	1	5	0	0	2	1	0	0	20	25.0000
TG	2	0	0	0	0	2	0	0	6	0	4	0	7	0	0	11	0	3	4	2	0	41	26.8293
TB	0	1	0	0	1	2	1	0	4	0	0	2	9	0	1	0	22	5	4	1	0	53	41.5094
TP	0	0	0	1	2	6	1	15	0	1	4	7	0	0	8	2	2	39	3	4	0	95	41.0526
TT	1	0	0	1	1	0	0	11	0	0	2	0	6	0	0	0	1	1	13	2	0	39	33.3333
XX	12	0	0	1	1	3	1	11	2	0	5	1	9	1	1	0	1	4	4	11	0	68	16.1765
Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	100.0000
Total	52	30	19	27	38	30	23	75	43	10	92	14	72	26	34	20	41	74	49	37	7	813	
User Accuracy(%)	36.5385	80.0000	100.0000	11.1111	57.8947	20.0000	47.8261	21.3333	27.9070	50.0000	63.0435	0.0000	23.6111	65.3846	14.7059	55.0000	53.6585	52.7027	26.5306	29.7297	71.4286		

5.2.2. Results obtained from RF

In RF, there are two major parameters which need to be defined before classification, namely, number of decision trees and the number of features per subset. The number of features per subset are taken as square root of the total number of features. In this study, there are 367 bands in data or 367 total number of features after the pre-processing step. Hence $\sqrt{367} \approx 19$ has been taken as number of features per subset. For studying the trend of variation of overall accuracy with the number of trees, a graph between the two has been plotted and shown in Figure 5.5. It can be observed that the maximum accuracy of 42.68% was obtained when number of trees were 300. Although 300 trees have given highest accuracy, the number of trees considered for training the classifier is 500 because error gets stabilized with 500 trees (Belgiu & Drăguţ, 2016). And the classifier can be compared properly by fixing this parameter.

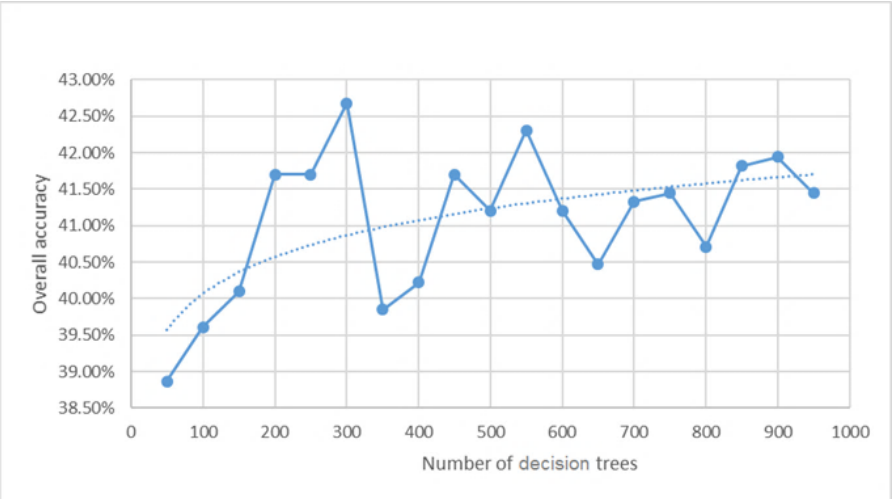
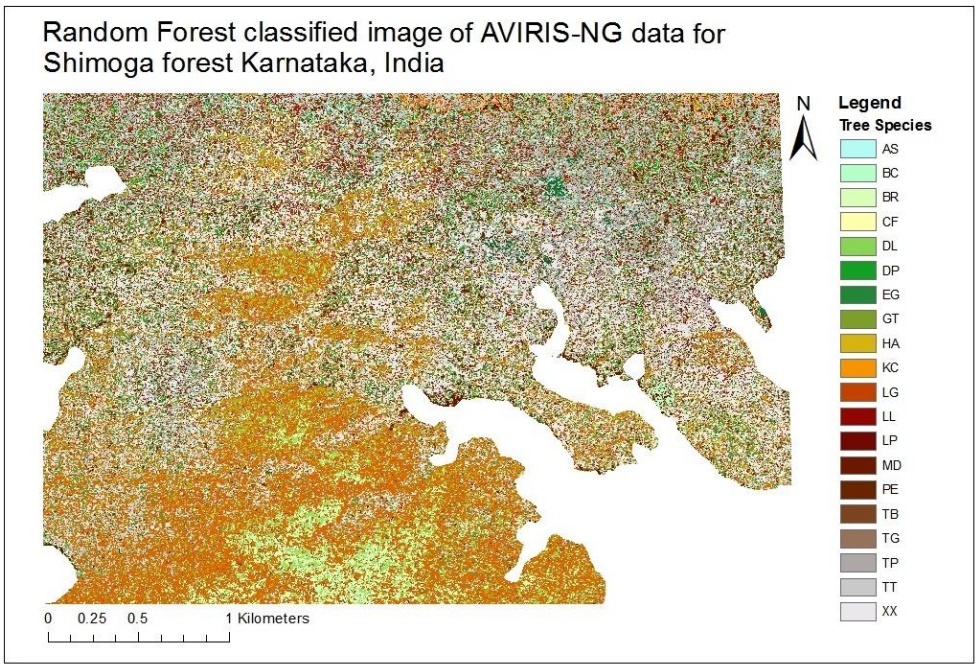


Figure 5.5: Variation in overall accuracy based on number of decision trees used in RF. Dotted line represents the trend of increase on logarithmic scale.

The classified image obtained after using RF with optimized parameters has been shown in Figure 5.6. From classified image, it can be observed that *Bahunia racemosa* is mostly confined to the southern part of the study area. *Bahunia racemosa*, *Dalbergia latifolia* and *Lannea grandis* are together forming a community in the southern part of the study area. *Lannea grandis* is the dominant species in the ridge of the Mandgadde range. In the northern part of the study area, many species were sparsely spread resulting in increasing species richness. No dominance of any particular species was seen. *Alsodephne semicarpifolia*, *Lagestormia Lanciolata*, *Terminalia paniculata* and *Tectona grandis* are mostly confined to the northern part only. *Terminalia paniculata* is more in north-eastern side, in the flat areas. *Terminalia paniculata* and *Xylia xylocarpa* were found to form a community. They were present together. *Lagestormia parviflora*, *Phyllanthus emblica*, *Bombax ceiba*, *Cassia fistula*, *Kydia calycina* and *Melia dubia,* all these classes were sparsely spread and no particular pattern can be observed. *Eucalyptus grandis* is present in clusters in only eastern part of the study area.



AS - Alsodephne semicarpifolia, BR - Bahunia racemosa, BC - Bombax ceiba, CF - Cassia fistula, DF - Dalbergia latifolia, DP - Delinia pentagyna, EG - Eucalyptus grandis, GT - Grewia tiliaefolia, HA - Holarrhena antidysenterica, KC - Kydia calycina, LL - Lagestormia lanciolata, LP - Lagestormia parviflora, LG - Lannea grandis, MD - Melia dubia PE - Phyllanthus emblica, TE - Tectona grandis, TB - Terminalia bellarica, TP - Terminalia paniculata, TT - Terminalia tomentosa, XX - Xylia xylocarpa

Figure 5.6: Classified output generated using Random Forest.

For accuracy assessment of RF, confusion matrix has been made as shown in Table 5.2. The RF has shown an overall accuracy of 40.34% and kappa value of 0.36. The OOB score obtained was 0.85. From the confusion matrix, it can be observed that class *Bombax ceiba* has shown maximum accuracy of 88.46% in which 23 testing pixels out of 26 were classified correctly. Second highest accuracy was observed in *Eucalyptus grandis* with 78.57% of accuracy where, out of 14 testing samples 11 were classified correctly. Followed by *Bahunia racemosa*, which has accuracy of 65.52% where out of 29 samples 19 were correctly classified. Lowest accuracy was observed for *Cassia fistula*, and it was highly misclassified with the one of the dominant species of the study area that is *Xylia xylocarpa*. In case of *Cassia fistula* only 2 pixels were correctly classified from among 33 pixels. Followed by *Alsodephne semicarpifolia* where only 5 pixels were correctly classified from 41 pixels with the producer accuracy of 12.19%. In case of *Xylia xylocarpa*, which is one of the dominant species of the study area. The classifier was trained with 104 pixels for this species. And this class was tested with 68 pixels. 35 pixels were correctly classified resulting in 51.47% of producer accuracy. But 86 pixels of other species has been misclassified with this class. Resulting in low user accuracy i.e. 22.73%.

Table 5.2: Confusion Matrix for RF

Ground reference point \ Clasified AVIRIS-NG Image	AS	BR	BC	CF	DL	DP	EG	GT	HA	KC	LL	LP	LG	MD	PE	TG	TB	TP	TT	XX	Other	Total	Producer Accuracy(%)
AS	5	0	0	0	1	2	0	3	1	0	2	0	0	0	5	5	0	5	5	7	0	41	12.1951
BR	0	19	0	0	2	0	0	0	1	0	0	0	6	0	0	0	1	0	0	0	0	29	65.5172
BC	0	0	23	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	26	88.4615
CF	1	0	0	2	0	0	0	2	0	0	0	0	0	0	1	0	0	5	0	22	0	33	6.0606
DL	0	13	0	0	22	0	0	0	3	0	0	0	5	0	0	0	1	0	0	0	0	44	50.0000
DP	0	0	0	1	0	7	0	1	0	1	1	0	2	1	0	0	0	0	3	3	0	20	35.0000
EG	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	1	0	2	0	14	78.5714
GT	1	0	1	0	1	1	0	11	0	4	3	0	8	4	1	2	0	2	5	9	0	53	20.7547
HA	1	0	0	0	0	0	0	0	13	0	1	0	2	0	0	1	0	0	0	6	0	24	54.1667
KC	0	0	0	0	0	3	0	0	0	8	0	0	1	5	0	0	0	0	4	0	0	21	38.0952
LL	3	0	0	4	0	0	0	18	2	0	34	1	3	0	6	2	0	4	1	16	0	94	36.1702
LP	0	0	0	4	2	0	0	3	0	2	0	4	0	0	0	0	2	0	2	11	0	30	13.3333
LG	0	0	0	1	0	0	0	6	1	0	0	0	18	0	1	0	0	0	1	9	0	37	48.6486
MD	0	0	4	1	0	0	0	0	0	1	0	0	0	12	2	0	1	1	0	4	0	26	46.1538
PE	1	0	0	0	0	1	0	3	0	0	1	0	1	0	6	0	0	1	1	5	0	20	30.0000
TG	1	0	0	3	0	4	0	1	1	0	7	0	4	0	0	14	0	2	2	2	0	41	34.1463
TB	0	1	0	0	0	0	0	0	8	0	0	1	11	0	3	0	14	4	4	7	0	53	26.4151
TP	2	0	1	0	1	0	2	7	0	0	7	4	1	1	7	0	1	48	2	11	0	95	50.5263
TT	0	0	0	2	0	0	0	7	0	0	3	0	5	0	0	0	0	0	17	5	0	39	43.5897
XX	6	0	1	3	0	2	0	3	3	0	2	0	4	0	3	0	0	4	2	35	0	68	51.4706
Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	100.0000
Total	21	33	30	21	29	20	13	65	33	16	63	10	71	23	35	24	20	78	49	154	5	813	
User (%) Accuracy	23.8095	57.5758	76.6667	9.5238	75.8621	35.0000	84.6154	16.9231	39.3939	50.0000	53.9683	40.0000	25.3521	52.1739	17.1429	58.3333	70.0000	61.5385	34.6939	22.7273	100.0000		

5.2.3. Result obtained from RoRF

The AVIRIS-NG hyperspectral imagery is classified using PCA based RoRF. For a better comparison with the RF classifier, the number of trees and the number of features per subset were considered to be same as RF i.e. 19 number of features per subset and 500 number of decision trees. The variation of number of decision trees with the overall accuracy is plotted in the graph shown in Figure 5.7. Maximum accuracy is seen when the decision trees were 400 to 450.

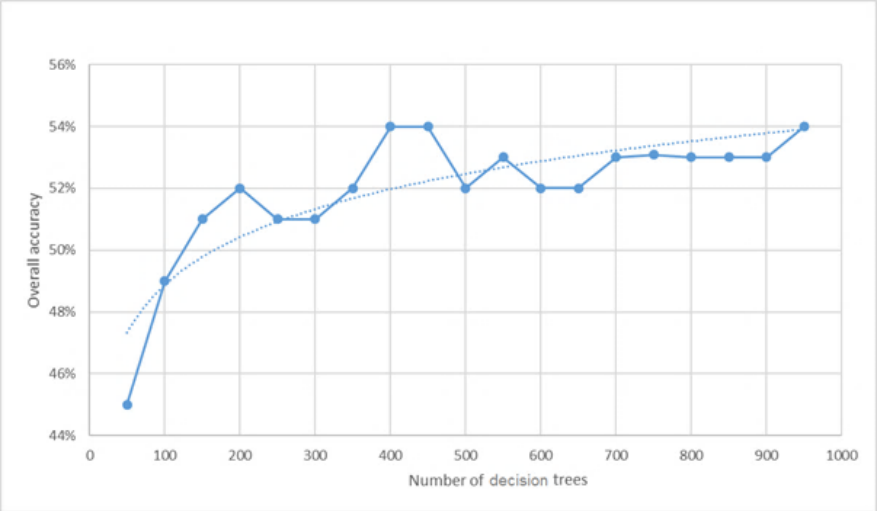


Figure 5.7: Variation in overall accuracy based on number of decision trees used in RoRF. Dotted line represents the trend of increase on logarithmic scale.

The classified image by RoRF is presented in Figure 5.8. The tree species *Lannea grandis* is spread throughout the ridge extended from north till south. With maximum spread in southern part of study area. Also, it is forming a community with *Bahunia racemosa* and *Dalbergia latifolia*. And confined to that area only. *Lannea grandis* is forming a community with *Holarrhena antidysenterica* in the ridge of the Mandgadde range. *Alsodephne semicarpifolia* is spread sparsely in the northern part of the study area. *Lagestormia Lanciolata* is also present in study area with *Alsodephne semicarpifolia* in the north side. But not forming any clear community together. *Terminalia paniculata* is present in all over the study area. It is sparsely spread in northern side but comparatively more in number. *Delinia pentagyna* is more in the eastern part of the study area. *Eucalyptus grandis* is present in clusters in the eastern side only. *Grewia tiliaefolia* is present in all over the study area except in the ridges present in the study area. *Lagestormia parviflora* is a rare species for the study area. It is found in very less number. *Tectona grandis* is sparsely present in the northern part of the study area. *Terminalia bellarica* is found in all over the study area. *Xylia xylocarpa* is one of the dominant species and is found more in flat areas of the study area in both eastern and western part of the study area. *Xylia xylocarpa*, *Terminalia paniculata* and *Terminalia tomentosa* are the three major species of this region. These three have formed a community. This community together dominates the eastern part of the study area.

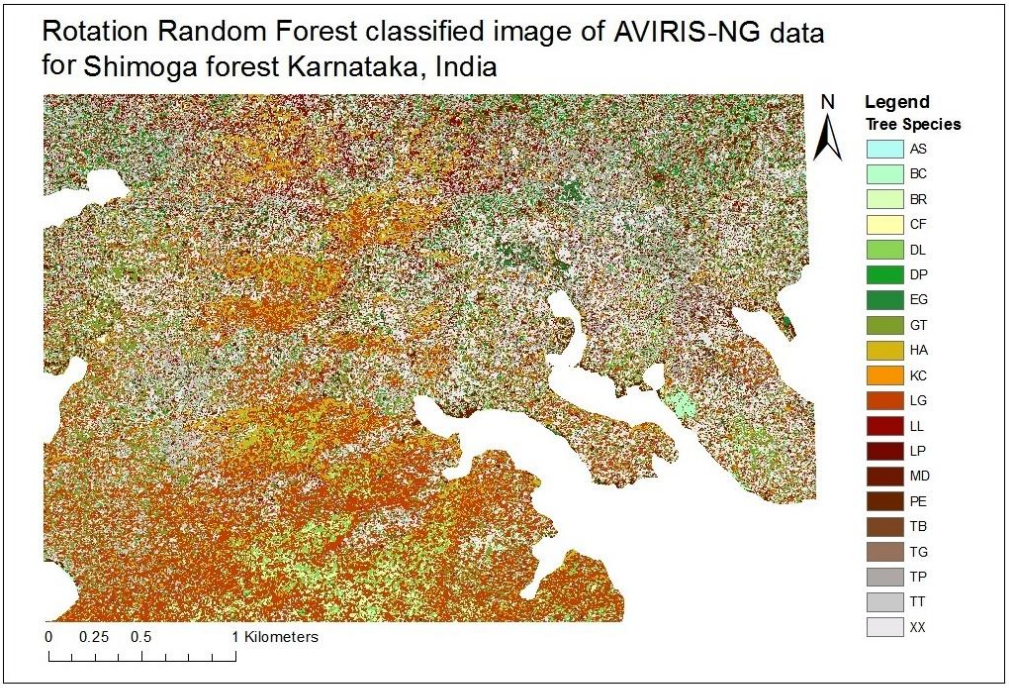


Figure 5.8: Classified output generated using Rotation Random Forest.

The accuracy assessment of the classifier is done by using confusion matrix given in Table 5.3. The RoRF gave an overall accuracy of 52.76% and kappa value of 0.49. The OOB score obtained was 0.85. RoRF has shown 100% accuracy in case of *Bombax ceiba* in which 17 samples have given to classifier for training and 26 samples were given for testing the classifier. RoRF has classified all the pixels correctly in case of *Bombax ceiba*. Second highest accuracy was seen in case of *Eucalyptus grandis* i.e. 85.71%, it has correctly classified 12 pixels among 14 pixels. *Dalbergia latifolia* and *Bahunia racemosa* has similar producer accuracy i.e. about 72%. In case of RoRF lowest accuracy was seen in class *Lagestormia parviflora*, which is a rare species in the study area. Out of 30 testing pixels only 5 were classified correctly. Most of them were misclassified with *Terminalia paniculata* and *Xylia xylocarpa*, *Lagestormia parviflora* is present as rare species in those area where *Xylia xylocarpa* and *Terminalia paniculata* are in majority and forming a community along with *Terminalia tomentosa*. For the dominant class *Xylia xylocarpa*, 39 out of 68 pixels were classified correctly while remaining pixels were misclassified with different species.

Table 5.3: Confusion matrix for RoRF

Ground reference point \ Classified AVIRIS-NG image	AS	BR	BC	CF	DL	DP	EG	GT	HA	KC	LL	LP	LG	MD	PE	TG	TB	TP	TT	XX	Other	Total	Producer Accuracy (%)
AS	13	0	0	0	1	2	0	3	1	0	5	0	2	0	4	0	0	4	2	4	0	41	31.7073
BR	0	21	0	0	2	0	0	0	1	0	0	0	5	0	0	0	0	0	0	0	0	29	72.4138
BC	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	100.0000
CF	1	0	0	10	0	0	0	2	0	0	0	0	3	0	2	0	0	3	1	11	0	33	30.3030
DL	0	3	0	0	32	0	3	0	4	0	0	0	2	0	0	0	0	0	0	0	0	44	72.7273
DP	1	0	0	1	0	10	0	0	0	0	1	0	1	1	0	0	0	1	2	2	0	20	50.0000
EG	0	0	0	0	0	0	12	0	0	0	0	0	0	0	1	0	0	0	0	1	0	14	85.7143
GT	0	0	0	0	1	4	0	18	0	1	3	0	6	3	2	1	0	2	5	7	0	53	33.9623
HA	1	0	0	0	0	0	0	2	14	0	1	0	2	0	0	0	0	0	0	4	0	24	58.3333
KC	0	0	0	0	0	2	0	2	0	8	0	0	0	6	0	0	0	0	3	0	0	21	38.0952
LL	1	0	0	4	0	0	0	7	1	1	61	0	5	0	1	3	0	2	1	7	0	94	64.8936
LP	0	0	0	1	2	0	0	0	0	0	0	5	2	0	1	0	1	9	0	9	0	30	16.6667
LG	0	0	0	0	0	0	0	4	1	0	0	0	21	1	0	0	0	2	1	7	0	37	56.7568
MD	0	0	0	0	0	1	0	3	0	1	0	0	0	18	0	0	0	0	0	3	0	26	69.2308
PE	0	0	0	0	0	0	0	4	0	0	0	0	1	0	10	0	0	0	2	3	0	20	50.0000
TG	0	0	0	3	0	3	0	0	4	0	7	0	6	0	0	11	0	4	2	1	0	41	26.8293
TB	0	0	0	1	4	2	0	0	4	0	0	3	12	0	2	0	15	4	2	4	0	53	28.3019
TP	0	0	0	1	1	1	0	5	0	0	6	3	2	0	4	0	2	60	2	8	0	95	63.1579
TT	0	0	0	0	0	0	0	7	0	0	3	0	3	0	0	0	1	0	20	5	0	39	51.2821
XX	4	0	0	0	0	2	0	5	3	0	4	0	4	1	1	0	0	4	1	39	0	68	57.3529
Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	100.0000
Total	21	24	26	21	43	27	15	62	33	11	91	11	77	30	28	15	19	95	44	115	5	813	
User (%) Accuracy	61.9048	87.5000	100.0000	47.6190	74.4186	37.0370	80.0000	29.0323	42.4242	72.7273	67.0330	45.4545	27.2727	60.0000	35.7143	73.3333	78.9474	63.1579	45.4545	33.9130	100.0000		

5.3. Comparison of classifiers

Producer accuracy is a measure of error of omission, which gives information to the producer of classification about how well a particular class can be classified (Congalton, 1991). So, it has given a special consideration in this study. Producer accuracy of all three classifiers with the training data for each class is given in the following graph shown in Figure 5.9. For *Eucalyptus grandis* and *Bahunia racemosa*, less training data was provided to the classifier even then all the three classifiers have given good accuracy. *Cassia fistula* and *Lagestormia parviflora* with 31 and 33 training samples respectively, producer accuracy was below 20% for all three classifiers.

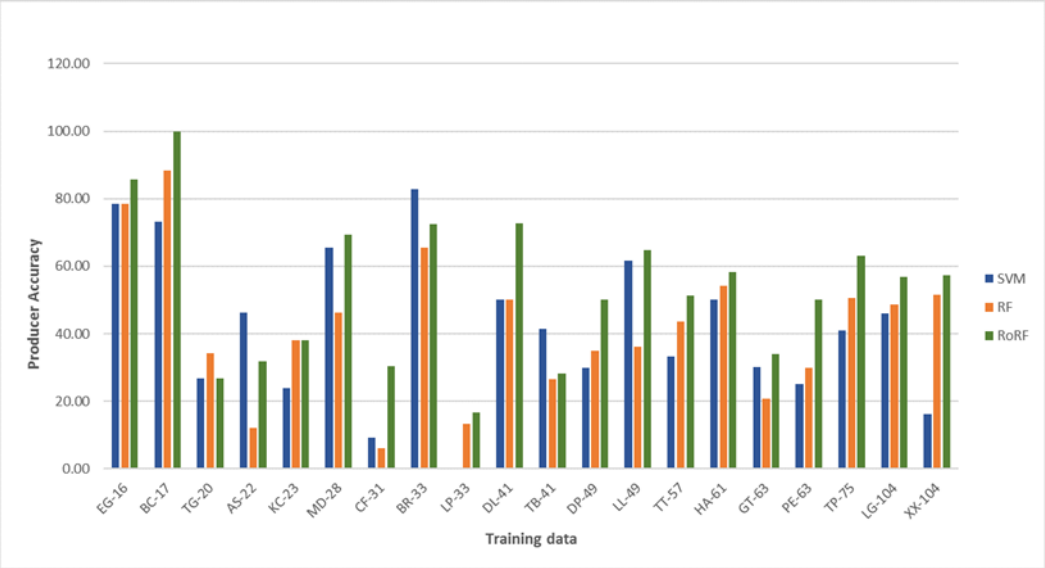


Figure 5.9: A comparison of producer accuracy for SVM, RF and RoRF along with the training data for each tree species class

Classifier comparison was also done by applying the McNemar test for checking whether the classification results are significantly different from each other or not. Following is the contingency matrix obtained by applying McNemar test. Table 5.4 shows the contingency matrix of the SVM and RF. Both of them classified 266 pixels correctly while 332 pixels were misclassified. In this case, z score obtained was 0.04. If z is between -1.96 and 1.96. The p value should be greater than 0.05. In this case p value is 0.08, hence accepted the null hypothesis that both the classifiers are similar in performance.

In case of RF and RoRF, the correctly classified pixels by both the classes were 276 while 345 were the misclassified pixels as shown in the Table 5.5. Note that these two values does not take part in the calculation of the test statistics. The number of pixels correctly classified by RF and wrongly classified by RoRF were 40 while the pixels correctly classified by RoRF and wrongly classified by RF were 137. These are the values which take part in calculating the test statistics. The z score obtained was 53.1582. Also, in case of SVM and RoRF as shown in the, the z score obtained was 44.1800. In both the cases z score was high and p value is very low as shown in Table 5.5 and Table 5.6 . The z score for all the three classifiers is shown in the following Table 5.7Table 5.7.

Table 5.4: Contingency matrix for McNemar test (RF and SVM)

	RF (wrong)	**RF (correct)**	**All**
SVM (wrong)	369	110	479
SVM (correct)	113	206	319
All	482	316	798
McNemar results	z score	0.0404	
	p value	0.8408	

Table 5.5: Contingency matrix for McNemar test (RF and RoRF)

	RF (wrong)	**RF (correct)**	**All**
RoRF (wrong)	345	40	385
RoRF (correct)	137	276	413
All	482	316	798
McNemar results	z score	53.1582	
	p value	$5.361 * 10^{-13}$	

Table 5.6: Contingency matrix for McNemar test (SVM and RoRF)

	RoRF (wrong)	**RoRF (correct)**	**All**
SVM (wrong)	332	147	479
SVM (correct)	53	266	319
All	385	413	798
McNemar results	z score	44.18	
	p value	$4.829 * 10^{-11}$	

Table 5.7: McNemar test results for SVM, RF and RoRF

	SVM	RF	RoRF
SVM	-	0.0404	44.1800
RF	-	-	53.1582
RoRF	-	-	-

5.4. Forest type map

With the help of working plan of the Shimoga region (Manjunath, 2011) trees identified in the field were labelled according to their type. In the study area, three types of forest vegetation was present; evergreen, deciduous and mix forest (both deciduous and evergreen species were present). Figure 5.10 shows the forest type map of Shimoga region. From map it can be seen that evergreen species are spread more in southern part and in the ridges of the hill in the centre part of the study area. Northern part of the study area is dominated by the deciduous species while eastern part and the flat areas of both east and west side of study area has mix type of forest vegetation. The Sentinel 2 imagery has also been analysed for the same represented in the Figure 3.2, In that it was clearly seen that the ridges of the Mandgadde Mountain range is dominated by evergreen species because even during the dry season the FCC is showing the presence of vegetation in that region. While in the other part of the region dominance of deciduous species can be clearly seen.

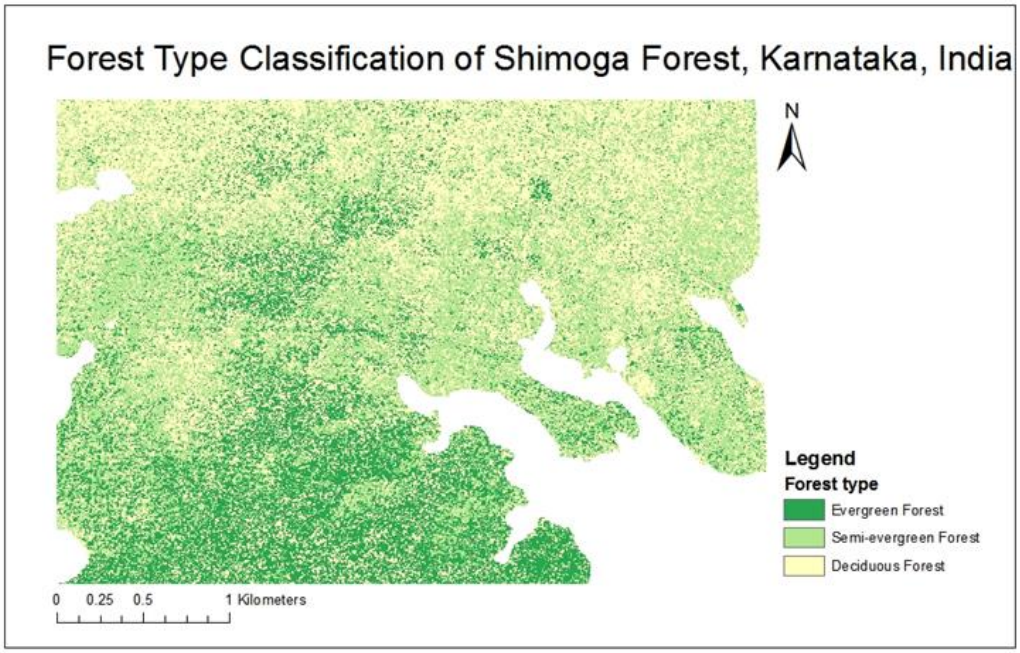


Figure 5.10: A forest type map of Shimoga Forest, Karnataka

5.5. Tree species richness

Species richness in the Shimoga region was estimated by using image classified with RoRF. Since the field plot size was 30m x 30m. So grid of same dimensions were selected to measure number of tree species in each grid. The variation in number of trees per 30m x 30m was found from 1 to 15 species, shown in Figure 5.11. Only in three grids, species richness was 1. Northern region has high species richness than the southern part of the study area. Flat areas has the maximum species richness in the eastern side.

Species richness from the RoRF classified map was compared with the collected field data. Following shows the location of the field plot with the species richness obtained from classified map and from collected ground data. Following graph in the Figure 5.12Figure 5.12 shows the clear comparison between them. Four plots has the same species richness from both field and from classified image. In case of three plots, species richness calculated from field is more than the classified map. Six plots has a variation of only one species and eight plots has the difference of only 2 species when compared richness from image with the richness from the field. Only in one plot difference in species richness was more i.e. 4 and three plots has difference of 3. These plots were mostly in the northern and eastern part of the study area as shown in Figure 5.13.

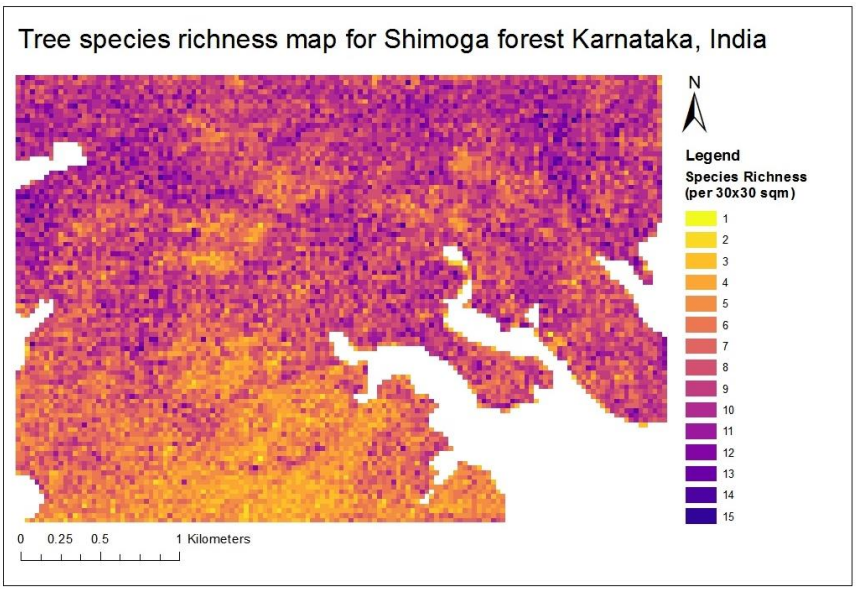


Figure 5.11: Tree species richness map for Shimoga Forest, Karnataka

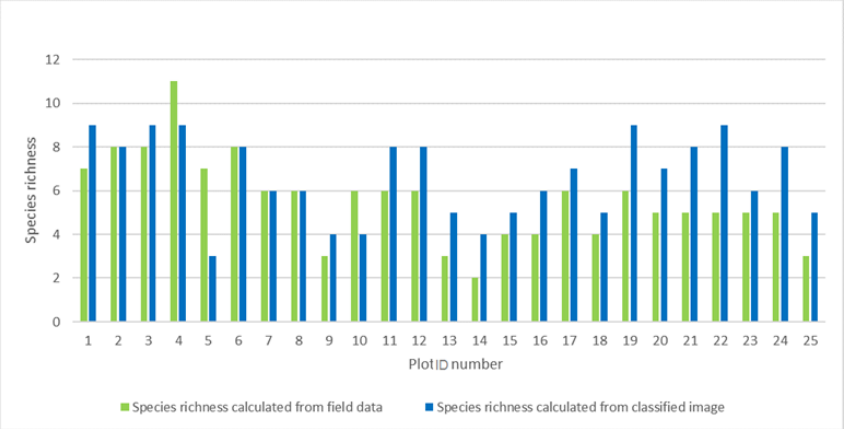


Figure 5.12: Comparison of tree species richness calculated from the field plot and the classified image

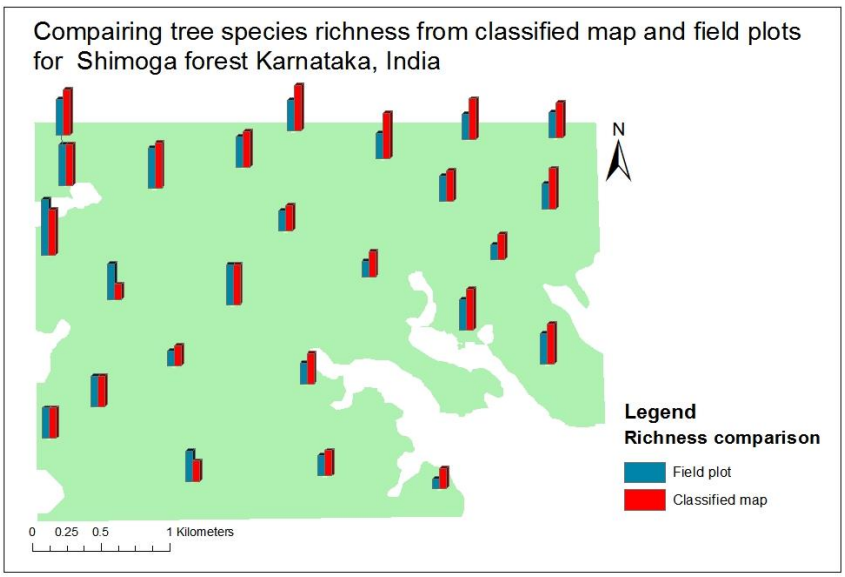


Figure 5.13: Comparison of tree species richness from classified map and field plots for Shimoga, Karnataka. Figure shows the species richness along with the location of plots

6. DISCUSSION

This chapter presents a discussion on the results reported in the Chapter 5. This includes the analysis done on the performance of the algorithms used in this study. This chapter also includes the discussion related to the forest tree species and about the species richness of the study area.

In this research, it can be seen that the PCA based transformed features at each node resulted in increasing the diversity of each individual classifier in the RoRF, which gave the decision trees with low correlation. Hence the accuracy got increased. If n features are chosen from the all N features, then $N + 1 - n$ will be the number of best-split feature, searched exhaustively. According to the splitting rule if we rank all the features, then the last $n - 1$ features will never get selected to split. According to the splitting rule, those feature which are in high rank with more discriminant power than other will have more chance to get selected. Hence, some feature will seldom be selected to split. In case of RF, since we have 367 number of features and 19 were selected for splitting, so at each node we can have at most 349 choices. But in case of RoRF, since each rotation matrix generates one unique split feature so, we can have 1.7310×10^{13} choices for the same (Zhang & Suganthan, 2014). Since the decision trees are very sensitive even to the small changes in the data so, more choices results in less correlated decision trees. Hence improving diversity among the decision trees.

All the three algorithms that were evaluated in this research were performed by considering all the features, without any feature selection from the whole dataset. This was done to have a fair comparison between their performances. To analyse the effect of feature selection in the algorithms performed in this study, feature selection was done using random forest and then all classifier were performed. Features were selected based on recursive eliminating variables method as proposed by Díaz-Uriarte & Alvarez de Andrés (2006). In this approach, first RF calculates the feature importance and then 20 % of the least important features were removed and then classification was performed on the remaining features. This was continuously performed by applying a loop till the accuracy started declining. As a result, in all three classifiers there was very less difference in the overall accuracy compared to the overall accuracy obtained when all the features were considered. The table showing the accuracy in case of all the three classifier is present in the Appendix C. This analysis was in consistent with the study done by Ferreira et al. (2016) for mapping tree species in tropical forest. They have mentioned that there used to be no significant difference in the accuracy after the feature selection.

In this research, the SVM has shown better performance in comparison of RF which was in agreement with the study done by Burai et al. (2015) and Raczko & Zagajewski (2017). They have also compared these classifiers for the species level classification and found that SVM outperforms RF. In this study, all the three classification results were tested by using McNemar test, where it was found that in the case of SVM and RF p value is 0.08, which was more than 0.05, hence the null hypothesis is accepted that both the classifiers are similar in performance. Hence SVM and RF do not significantly differ from each other. This result is in concordance with the study done by Ghosh et al. (2014), they have also found that for hyperspectral data, SVM and RF shows the similar performance. When SVM and RF is compared with RoRF the p value is less than 0.05, hence the null hypothesis is rejected that means the classification results are significantly different. Hence the performance of RoRF is significantly different from both SVM and RF which is in consistence with the study done by Zhang & Suganthan (2014), where PCA based RoRF has outperformed RF.

In case of almost all the species classes, RoRF have shown good accuracy than the SVM and RF. The reason behind this was that in case of hyperspectral data where large number of features are present, although RF performs well. But in case of RoRF, more randomization is included during selection of feature subset in the RoRF, classifiers become diverse and as a result accuracy increases. It can be noticed that the species are misclassified with the dominant species of the study area. *Eucalyptus grandis* has shown good accuracy in all three cases. This species was present in the patches. Since *Eucalyptus grandis* shows allelopathic behaviour, i.e. it releases chemical which inhibits other species around them. Hence it was present in separate groups so a large number of pure pixels were obtained for the species. *Lagestormia parviflora* has shown the accuracy less than 20 % in all three classifiers due of unavailability of the good quality training data. *Lagestormia parviflora* in group were not found anywhere in the 25 plots. To have better classification of the species, training data should be collected from the areas where the tree is present in groups. This supports the fact that for hyperspectral data, it is better to collect data from the region where the group of same trees are available. This is done to reduce the variability between the same species in different locations which comes due to the crown mixing and background signals (Carleer & Wolff, 2004; Ghosh et al., 2014). In case of tree class *Bombax ceiba,* it is a very large tree, which flowers from December to February. It has a large (approx. 10cm across), dark crimson coloured flower due to which, its spectral signatures has high variability than others. Hence it resulted in attaining the high producer accuracy by each classifier. *Lannea grandis* showed lower user accuracy because many other species has been misclassified with it. This happened because variability in the spectral signatures of *Lannea grandis* is very low. Hence, Species with high variance in their spectral signature showed good accuracy in comparison to other species with low spectral variability.

In the eastern side of the forest type map, the area is majorly dominated by more number of deciduous and semi-evergreen species. While in the area having the ridges is dominated by evergreen species. According to the Shimoga Forest working plan (Manjunath, 2011), in the eastern part of the study area past illegal felling and forest fire is reported. Due to which blank patches were left at this region. So forest department did plantation to fill those blank patches inside the natural forest. As a result, mix vegetation has resulted including both deciduous and evergreen species. The species richness from classified image and from the field data is same for some field plots but for some plots it varies. The reason behind the variation of species richness is due to the fact that the spatial resolution of the hyperspectral imagery is 5m x 5m. While there were only a very few cases where the span of 25sq. m was being occupied by single tree species on the field. This reduced the amount of pure pixels for classification. Cases were observed where one pixel was occupied by more than one tree crown of the training data. In such cases, the pixel was labelled as the species whose crown covered the centre point of the pixel. This mislabelling further lead to misclassification and hence cause the variation in the species richness as well.

The Figure 4.4 presented in section 4.2 represents the TWI of the study area which has been calculated by using DEM of 30 m. Since the lower value of the TWI represents ridges while the higher values represents drainage depressions (Cooley, 2015). The areas where TWI was low, the richness was also found to be low. While in the eastern and western part of the study area has high species richness, it was because of the high drainage depressions present over there. Species richness is more where TWI is more. This result is in agreement with the fact that both TWI and species richness are positively correlated shown in the study done by Song & Cao (2017). In the southern part of the study area, species richness is low while in the northern and the eastern part, species richness is very high, because in the past illegal felling and forest fire was reported in that area. So plantation with different species was done on this region. Exotic species like *Tectona grandis* and *Eucalyptus grandis* were also found in this area. Hence resulting in more number of different species.

7. CONCLUSION AND RECOMMENDATIONS

7.1. Conclusion

This research has shown the performance of three machine learning algorithms that are Support Vector Machine (SVM), Random Forest (RF) and PCA based Rotation Random Forest (RoRF) for mapping of tree species richness of tropical forest of Shimoga, Karnataka. This was achieved by classifying the imagery obtained by AVIRIS NG hyperspectral sensor. PCA based RoRF model was proposed in this study to improve the diversity of individual classifiers in the ensemble. Which ultimately results in increasing the accuracy of the classifier. In this method the feature set is split into k subsets. PCA is applied separately in each subset to transform data into another space and all the components were kept, resulting in a new extracted feature set. With this new extracted feature dataset, decision trees are trained. And each rotation generates a unique split. Thus resulting into diverse classifiers.

The conclusion that can be drawn from this research is that the RoRF has outperformed the other two algorithms. McNemar test was performed to evaluate the performance of the classifiers. The test suggests that the performance of RoRF is significantly different from SVM and RF. Also, there is no significant difference in the performance of the SVM and RF. To achieve the objectives, the result obtained from the best performed algorithm i.e. RoRF was then used for the estimation of the tree species richness in the study area. The plot wise species richness is calculated manually i.e. by counting the unique species in each plot. The obtained tree species richness from the RoRF classified map is then compared with the data collected from the field survey. As a result it is found that, out of 25 plots, in some of the plots, species richness was same as that of the richness obtained from the imagery while in some plots species richness was slightly different. Only in some cases the difference was more.

7.2. Answers to research questions

This study answers the following research questions that were proposed:

Question: How does the PCA-based Rotation Random Forest classifier perform in comparison with RFs and SVM?
Answer: From the results obtained in section 5.3, it can be inferred that PCA based RoRF has significantly outperformed the other two classifiers, SVM and RF. This indicates that RoRF can better deal with the high dimensional feature space and the limited number of training samples. The reason for this is that more diversity and higher accuracy has been introduced in the individual classifier of the ensemble.

Question: How to deal with the unclassified tree species classes that are present in the study area?
Answer: A total of 20 unique tree species samples were collected from the study area. This study made an assumption that the study area consists of only these 20 tree species, and the remaining existent tree species were ignored due to unavailability of appropriate training data. Although, on comparison with the working plan of Shimoga (Manjunath, 2011), it was noted that most of the dominant species of the region were included in the study. Also, for dealing with remaining unclassified classes, a masking region was defined to clip out the region.

Question: To what extent can we map tree species from the used hyperspectral data?
Answer: Hyperspectral data, due to its very high spectral resolution has the ability to capture the highly detailed information relating to individual species, which makes it easier to uniquely identify each tree species which can be seen in section 5.1. In the tropical forest, mapping of tree species is very difficult but in this study, the hyperspectral data was able to map 20 species of the tropical forest. Those species which were not available in groups showed less accuracy. But the species available in group showed good accuracy. It can be interpreted that with the quality training data, species mapping can be improved further by using hyperspectral data.

Question: To what extent the species richness obtained from classified image differs from the species richness obtained from field data?
Answer: From the results of section 5.5, it can be seen that a mix variation of species richness was observed when the classified tree species richness was compared with field tree species richness. There were cases where the species richness obtained from classified image was same as species richness obtained from the field. While in some cases, there was a variation which ranged from difference of count of 1 species up to 4 species. The reason behind this variation was the spatial resolution of the data used. The area covered by one individual pixel was 25 sq. m while this was capable of covering more than a single tree species, which lead to unavailability of pure pixels for the training of the classifier. This reason caused the classifier to misclassify some of the classes and hence causing the variation in the tree species richness.

7.3. Recommendations

The future scope of this research has been highlighted in form of following recommendations:

- In this research, location information of tree species was collected as training and testing samples for classifiers and spectral signature generation to see the variability among them. So, use of GPS with high accuracy or Differential Global Positioning System (DGPS) can be used for ground data collection in conjunction with the location information.
- In this study RoRF was compared with only two classifiers, i.e. SVM and RF, there is still scope to explore its potential by comparing it with other classifiers. Also, different type of base classifier can be explored to undermine the best suited base classifier for RoRF.
- This research makes use of PCA for the purpose of feature extraction. This opens up scope to explore other feature extraction techniques and compare their performance.
- Ancillary data like topography and DEM can be used in tree species classification process for improved results. Other sensor data like LiDAR can be combined to improve the accuracy as mentioned in some studies (Ghosh et al., 2014; Leutner et al., 2012). The use of LiDAR data will enrich the information of the tree species by capturing the height and the canopy of the trees.

LIST OF REFERENCES

Akbari, H., & Kalbi, S. (2017). Determining Pleiades satellite data capability for tree diversity modeling, *10*(iForest 10), 348–352. https://doi.org/10.3832/ifor1884-009

Bahria, S., Essoussi, N., & Limam, M. (2011). Hyperspectral data classification using geostatistics and support vector machines. *Remote Sensing Letters*, *2*(2), 99–106. https://doi.org/10.1080/01431161.2010.497782

Baldeck, C. A., & Asner, G. P. (2014). Improving Remote Species Identification through Efficient Training Data Collection. *Remote Sensing*, *6*(4), 2682–2698. https://doi.org/10.3390/rs6042682

Baldeck, C. A., & Asner, G. P. (2015). Single-Species Detection With Airborne Imaging Spectroscopy Data: A Comparison of Support Vector Techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(6), 2501–2512. https://doi.org/10.1109/JSTARS.2014.2346475

Baldeck, C. A., Asner, G. P., Martin, R. E., Anderson, C. B., Knapp, D. E., Kellner, J. R., & Wright, S. J. (2015). Operational Tree Species Mapping in a Diverse Tropical Forest with Airborne Imaging Spectroscopy. *PLOS ONE*, *10*(7), e0118403. https://doi.org/10.1371/journal.pone.0118403

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

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, *24*(1), 43–69. https://doi.org/10.1080/02626667909491834

Bhat, S., Chandran, M. D., & Ramachandra, T. V. (2012). Status of Forests in Shimoga, Central Western Ghats. In *Proceedings of the LAKE 2012: National Conference on Conservation and Management of Wetland Ecosystems, 06th - 09th November 2012* (pp. 1–10). School of Environmental Sciences, Mahatma Gandhi University, Kottayam, Kerala. Retrieved from http://wgbis.ces.iisc.ernet.in/energy/water/paper/lake2012_forest_shimoga/index.htm

Bostanci, B., & Bostanci, E. (2013). An Evaluation of Classification Algorithms Using Mc Nemar's Test (pp. 15–26). Springer, India. https://doi.org/10.1007/978-81-322-1038-2_2

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burai, P., Deák, B., Valkó, O., Tomor, T., Burai, P., Deák, B., … Tomor, T. (2015). Classification of Herbaceous Vegetation Using Airborne Hyperspectral Imagery. *Remote Sensing*, *7*(2), 2046–2066. https://doi.org/10.3390/rs70202046

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Burkmar, R. (2015). Tom.bio QGIS Biological Recording Plugin | Biodiversity Projects. Retrieved February 18, 2019, from https://www.fscbiodiversity.uk/?q=qgis-plugin

Carleer, A., & Wolff, E. (2004). Exploitation of Very High Resolution Satellite Data for Tree Species Identification. *Photogrammetric Engineering & Remote Sensing*, *70*(1), 135–140. https://doi.org/10.14358/PERS.70.1.135

Carlson, K. M., Asner, G. P., Hughes, R. F., Ostertag, R., & Martin, R. E. (2007). Hyperspectral Remote Sensing of Canopy Biodiversity in Hawaiian Lowland Rainforests. *Ecosystems*, *10*(4), 536–549. https://doi.org/10.1007/s10021-007-9041-z

Champion, H. G., & Seth, S. K. (1968). A revised survey of the forest types of India. New Delhi: Manager of Publications. Retrieved from https://dds.crl.edu/crldelivery/23005

Chutia, D., Bhattacharyya, D. K., Sarma, K. K., Kalita, R., & Sudhakar, S. (2016). Hyperspectral Remote Sensing Classifications: A Perspective Survey. *Transactions in GIS*, *20*(4), 463–490. https://doi.org/10.1111/tgis.12164

Clark, M. L., & Roberts, D. A. (2012). Species-Level Differences in Hyperspectral Metrics among Tropical Rainforest Trees as Determined by a Tree-Based Classifier. *Remote Sensing*, *4*(6), 1820–1855.

https://doi.org/10.3390/rs4061820
Clark, M. L., Roberts, D. A., & Clark, D. (2005). Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote Sensing of Environment*, *96*(3–4), 375–398. https://doi.org/10.1016/j.rse.2005.03.009
Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, *37*(1), 35–46. https://doi.org/10.1016/0034-4257(91)90048-B
Cooley, S. W. (2015). Topographic Wetness Index (TWI) | GIS 4 Geomorphology. Retrieved December 8, 2018, from http://gis4geomorphology.com/topographic-index-model/
Cortés Rodríguez, H. (2014). Classifiers ensemble in remote sensing: a comparative analysis. Retrieved from http://repositori.uji.es/xmlui/handle/10234/158826
Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3. https://doi.org/10.1186/1471-2105-7-3
Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197
Duro, D. C., Coops, N. C., Wulder, M. A., & Han, T. (2007). Development of a large area biodiversity monitoring system driven by remote sensing. *Progress in Physical Geography*, *31*(3), 235–260. https://doi.org/10.1177/0309133307079054
Engler, R., Waser, L. T., Zimmermann, N. E., Schaub, M., Berdos, S., Ginzler, C., & Psomas, A. (2013). Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. *Forest Ecology and Management*, *310*, 64–73. https://doi.org/10.1016/J.FORECO.2013.07.059
Fassnacht, F. E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L. T., … Ghosh, A. (2016). Review of studies on tree species classification from remotely sensed data. *Remote Sensing of Environment*, *186*, 64–87. https://doi.org/10.1016/J.RSE.2016.08.013
Felbermeier, B., Hahn, A., & Schneider, T. (2010). *Study on user requirements for remote sensing applications in forestry*. Retrieved from http://www.isprs.org/proceedings/xxxviii/part7/b/pdf/210_XXXVIII-part7B.pdf
Feret, J.-B., & Asner, G. P. (2012). Semi-supervised methods to identify individual crowns of lowland tropical canopy species using imaging spectroscopy and LiDAR. *Remote Sens*, *4*.
Feret, J.-B., & Asner, G. P. (2013). Tree Species Discrimination in Tropical Forests Using Airborne Imaging Spectroscopy. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(1), 73–84. https://doi.org/10.1109/TGRS.2012.2199323
Ferreira, M. P., Zortea, M., Zanotta, D. C., Shimabukuro, Y. E., & Filho, C. R. de S. (2016). Mapping tree species in tropical seasonal semi-deciduous forests with hyperspectral and multispectral data. *Remote Sensing of Environment*, *179*, 66–78. https://doi.org/10.1016/j.rse.2016.03.021
Foody, G. M., & Cutler, M. E. J. (2003). Tree biodiversity in protected and logged Bornean tropical rain forests and its measurement by satellite remote sensing. *Journal of Biogeography*, *30*(7), 1053–1066. https://doi.org/10.1046/j.1365-2699.2003.00887.x
Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/JCSS.1997.1504
Fricker, A. (2017). Installing the Topographic Wetness Index Python Script into an ArcGIS Toolbox - YouTube. Retrieved December 9, 2018, from https://www.youtube.com/watch?v=6g75Ce3ZQwQ&t=138s
George, R., Padalia, H., & Kushwaha, S. P. S. (2014). Forest tree species discrimination in western Himalaya using EO-1 Hyperion. *International Journal of Applied Earth Observation and Geoinformation*, *28*(1), 140–149. https://doi.org/10.1016/j.jag.2013.11.011
Ghosh, A., Fassnacht, F. E., Joshi, P. K., & Koch, B. (2014). A framework for mapping tree species combining hyperspectral and LiDAR data: Role of selected classifiers and sensor across three spatial scales. *International Journal of Applied Earth Observation and Geoinformation*, *26*, 49–63. https://doi.org/10.1016/J.JAG.2013.05.017
gisresources. (2013). Fundamemtals of Hyperspectral Remote Sensing - GIS Resources. Retrieved March 4, 2019, from http://www.gisresources.com/fundamemtals-of-hyperspectral-remote-sensing_2/
Goetz, A. F., Vane, G., Solomon, J. E., & Rock, B. N. (1985). Imaging spectrometry for Earth remote sensing. *Science (New York, N.Y.)*, *228*(4704), 1147–1153. https://doi.org/10.1126/science.228.4704.1147
Google Earth. (2019). Retrieved March 3, 2019, from http://www.google.com/earth/download/ge/

Heijden, F. van der., Duin, R. P. W., Ridder, D. de, & Tax, D. M. J. (2004). *Classification, parameter estimation, and state estimation : an engineering approach using MATLAB*. Wiley. Retrieved from https://research.utwente.nl/en/publications/classification-parameter-estimation-and-state-estimation-an-engin

Jaramillo, F., & Destouni, G. (2015). Comment on "planetary boundaries: Guiding human development on a changing planet." *Science*, *348*(6240), 1217–c. https://doi.org/10.1126/science.aaa9629

Jenkins, B. (2015). Latin hypercubes and all that: How DOE works | Ora Research. Retrieved February 17, 2019, from http://oraresearch.com/2015/02/latin-hypercubes-and-all-that/

Jensen, R. R., Hardin, P. J., & Hardin, A. J. (2012). Classification of urban tree species using hyperspectral imagery. *Geocarto International*, *27*(5), 443–458. https://doi.org/10.1080/10106049.2011.638989

Jia, X., & Richards, J. A. (1994). Efficient maximum likelihood classification for imaging spectrometer data sets. *IEEE Transactions on Geoscience and Remote Sensing*, *32*(2), 274–281. https://doi.org/10.1109/36.295042

Kalacska, M., Sanchez-Azofeifa, G. A., Rivard, B., Caelli, T., White, H. P., & Calvo-Alvarado, J. C. (2007). Ecological fingerprinting of ecosystem succession: Estimating secondary tropical dry forest structure and diversity using imaging spectroscopy. *Remote Sensing of Environment*, *108*(1), 82–96. https://doi.org/10.1016/J.RSE.2006.11.007

Korpela, I., Hovi, A., & Morsdorf, F. (2012). Understory trees in airborne LiDAR data — Selective mapping due to transmission losses and echo-triggering mechanisms. *Remote Sensing of Environment*, *119*, 92–104. https://doi.org/10.1016/J.RSE.2011.12.011

Laurin, G. V., Chan, J. C.-W., Chen, Q., Lindsell, J. A., Coomes, D. A., Guerriero, L., … Valentini, R. (2014). Biodiversity Mapping in a Tropical West African Forest with Airborne Hyperspectral Data. *PLoS ONE*, *9*(6), e97910. https://doi.org/10.1371/journal.pone.0097910

Leckie, D. G., Tinis, S., Nelson, T., Burnett, C., Gougeon, F. A., Cloney, E., & Paradine, D. (2005). Issues in species classification of trees in old growth conifer stands. *Canadian Journal of Remote Sensing*, *31*(2), 175–190. https://doi.org/10.5589/m05-004

Leutner, B. F., Reineking, B., Müller, J., Bachmann, M., Beierkuhnlein, C., Dech, S., … Wegmann, M. (2012). Modelling Forest α-Diversity and Floristic Composition — On the Added Value of LiDAR plus Hyperspectral Remote Sensing. *Remote Sensing*, *4*(9), 2818–2845. https://doi.org/10.3390/rs4092818

Liu, B., Dai, Y., Li, X., Lee, W. S., & Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining* (pp. 179–186). IEEE Comput. Soc. https://doi.org/10.1109/ICDM.2003.1250918

M. Foody, G., Atkinson, P., Gething, P., Ravenhill, N., & Kelly, C. (2005). Identification of specific tree species in ancient semi-natural woodland from digital aerial sensor imagery. *Ecological Applications - ECOL APPL*, *15*, 1233–1244.

Madonsela, S., Cho, M. A., Ramoelo, A., Mutanga, O., & Naidoo, L. (2018). Estimating tree species diversity in the savannah using NDVI and woody canopy cover. *International Journal of Applied Earth Observation and Geoinformation*, *66*, 106–115. https://doi.org/10.1016/j.jag.2017.11.005

Magurran, A. E. (2004). *Measuring biological diversity*. Blackwell Pub. Retrieved from https://www.wiley.com/en-us/Measuring+Biological+Diversity-p-9780632056330

Maity, S., Patnaik, C., Das, A., & Misra, A. (2017). *Absorption Peak Decomposition Technique for Forest. Species Identification from AVIRIS-NG Hyperspectral Data - Science Results from Phase – 1 Airborne Hyperspectral Campaign with AVIRIS-NG over India*. Bangalore. Retrieved from https://vedas.sac.gov.in/aviris/pdf/AVIRIS_NG_FIRST_PHASE_REPORT.pdf

Manjunath, K. B. (2011). *Working plan for the forests of Bengaluru rural division (2011-12 to 2020-21)*. Bengaluru. Retrieved from http://aranya.gov.in/new/newdownloads/WP/Bangalore Rural WP.pdf

Mather, P. M., & Koch, M. (2011). *Computer Processing of Remotely-Sensed Images*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470666517

McGwire, K., Minor, T., & Fenstermaker, L. (2000). Hyperspectral Mixture Modeling for Quantifying Sparse Vegetation Cover in Arid Environments. *Remote Sensing of Environment*, *72*(3), 360–374. https://doi.org/10.1016/S0034-4257(99)00112-1

Millennium Ecosystem Assessment. (2005). *Ecosystems and Human Well-being: Synthesis*. Washington, DC. Retrieved from https://www.millenniumassessment.org/documents/document.356.aspx.pdf

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, *32*(9), 1378–1388. https://doi.org/10.1016/J.CAGEO.2005.12.009

Moeslund, J. E., Arge, L., Bøcher, P. K., Dalgaard, T., & Svenning, J.-C. (2013). Topography as a driver of local terrestrial vascular plant diversity patterns. *Nordic Journal of Botany*, *31*(2), 129–144. https://doi.org/10.1111/j.1756-1051.2013.00082.x

Palmer, M. W., Earls, P. G., Hoagland, B. W., White, P. S., & Wohlgemuth, T. (2002). Quantitative tools for perfecting species lists. *Environmetrics*, *13*(2), 121–137. https://doi.org/10.1002/env.516

Pandey, P. C., Tate, N. J., & Balzter, H. (2014). Mapping Tree Species in Coastal Portugal Using Statistically Segmented Principal Component Analysis and Other Methods. *IEEE SENSORS JOURNAL*, *14*(12). https://doi.org/10.1109/JSEN.2014.2335612

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825--2830. Retrieved from https://scikit-learn.org/stable/about.html#id1

Pouteau, R., Gillespie, T., & Birnbaum, P. (2018). Predicting Tropical Tree Species Richness from Normalized Difference Vegetation Index Time Series: The Devil Is Perhaps Not in the Detail. *Remote Sensing*, *10*(5), 698. https://doi.org/10.3390/rs10050698

Purvis, A., & Hector, A. (2000). Getting the Measure of Biodiversity. *Nature*, *405*, 212–219.

QGIS Development Team. (2009). QGIS Geographic Information System. *Open Source Geospatial Foundation*. Retrieved from http://qgis.org

R Development Core Team. (2010). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org/

Raczko, E., & Zagajewski, B. (2017). Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. *European Journal of Remote Sensing*, *50*(1), 144–154. https://doi.org/10.1080/22797254.2017.1299557

Richards, J. A., & Jia, X. (2006). *Remote sensing digital image analysis : an introduction*. Springer.

Roberts, D. A., Gardner, M., Church, R., Ustin, S., Scheer, G., & Green, R. O. (1998). Mapping Chaparral in the Santa Monica Mountains Using Multiple Endmember Spectral Mixture Models. *Remote Sensing of Environment*, *65*(3), 267–279. https://doi.org/10.1016/S0034-4257(98)00037-6

Robichaud, P. R., Lewis, S. A., Laes, D. Y. M., Hudak, A. T., Kokaly, R. F., & Zamudio, J. A. (2007). Postfire soil burn severity mapping with hyperspectral image unmixing. *Remote Sensing of the Environment. 108: 467-480.*, 467–480. Retrieved from https://www.fs.usda.gov/treesearch/pubs/29030

Rocchini, D. (2007). Effects of spatial and spectral resolution in estimating ecosystem α-diversity by satellite imagery. *Remote Sensing of Environment*, *111*(4), 423–434. https://doi.org/10.1016/J.RSE.2007.03.018

Rocchini, D., Ricotta, C., & Chiarucci, A. (2009). Using satellite imagery to assess plant species richness: The role of multispectral systems. *Applied Vegetation Science*, *10*(3), 325–331. https://doi.org/10.1111/j.1654-109X.2007.tb00431.x

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S. I., Lambin, E., … Foley, J. (2009). Planetary Boundaries: Exploring the Safe Operating Space for Humanity. *Ecology and Society*, *14*(2), art32. https://doi.org/10.5751/ES-03180-140232

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1619–1630. https://doi.org/10.1109/TPAMI.2006.211

Roger, R. E. (1996). Sparse inverse covariance matrices and efficient maximum likelihood classification of hyperspectral data. *International Journal of Remote Sensing*, *17*(3), 589–613. https://doi.org/10.1080/01431169608949029

Roy, P. S., Behera, M. D., Murthy, M. S. R., Roy, A., Singh, S., Kushwaha, S. P. S., … Ramachandran, R. M. (2015). New vegetation type map of India prepared using satellite remote sensing: Comparison with global vegetation maps and utilities. *International Journal of Applied Earth Observation and Geoinformation*, *39*, 142–159. https://doi.org/10.1016/j.jag.2015.03.003

Skurichina, M., & Duin, R. P. W. (2005). Combining Feature Subsets in Feature Selection (pp. 165–175). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11494683_17

Sobhan, M. I. (2007). *Species discrimination from a hyperspectral perspective*. Enschede. Retrieved from http://edepot.wur.nl/121927

Song, C., & Cao, M. (2017). Relationships between Plant Species Richness and Terrain in Middle Sub-Tropical Eastern China. *Forests*, *8*(9), 344. https://doi.org/10.3390/f8090344

Stockholm Resilience Centre. (2012). Planetary boundaries - Stockholm Resilience Centre. Retrieved June 6, 2018, from http://www.stockholmresilience.org/research/planetary-boundaries.html

The Times of India. (2012). UN designates Western Ghats as world heritage site - Times of India. Retrieved March 4, 2019, from https://timesofindia.indiatimes.com/home/environment/flora-fauna/UN-designates-Western-Ghats-as-world-heritage-site/articleshow/14595602.cms?referral=PM

Tso, B., & Mather, P. (2009). *Classification Methods for Remotely Sensed Data, Second Edition*. CRC Press. https://doi.org/10.1201/9781420090741

Tumer, K., & Oza, N. C. (2003). Input decimated ensembles. *Pattern Analysis & Applications*, *6*(1), 65–77. https://doi.org/10.1007/s10044-002-0181-7

Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E., & Steininger, M. (2003). Remote sensing for biodiversity science and conservation. *Trends in Ecology & Evolution*, *18*(6), 306–314. https://doi.org/10.1016/S0169-5347(03)00070-3

UNDP. (2016). Goal 15 targets | UNDP. Retrieved June 6, 2018, from http://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-15-life-on-land/targets/

van Aardt, J. A. N., & Wynne, R. H. (2007). Examining pine spectral separability using hyperspectral data from an airborne sensor: An extension of field-based results. *International Journal of Remote Sensing*, *28*(2), 431–436. https://doi.org/10.1080/01431160500444772

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Berlin, Heidelberg: Springer-Verlag.

Vedas SAC. (2016). AVIRIS-NG Sites Visualization and Data Download. Retrieved from https://vedas.sac.gov.in/aviris_2.0/sitemap.html

Wang, L., Sousa, W. P., & Gong, P. (2004). Integration of object-based and pixel-based classification for mapping mangroves with IKONOS imagery. *International Journal of Remote Sensing*, *25*(24), 5655–5668. https://doi.org/10.1080/01431160233129121 5

Wang, L., & Zhao, C. (2016). *Hyperspectral Image Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-47456-3

Webb, A. R. (2002). *Statistical Pattern Recognition*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/0470854774

Wei, T., & Simko, V. (2017). R package "corrplot": Visualization of a Correlation Matrix. Retrieved March 3, 2019, from https://github.com/taiyun/corrplot

Williams, A. P., & Hunt, E. R. (2002). Estimation of leafy spurge cover from hyperspectral imagery using mixture tuned matched filtering. *Remote Sensing of Environment*, *82*(2–3), 446–456. https://doi.org/10.1016/S0034-4257(02)00061-5

Wilson, E. O. (1993). *The diversity of life*. W.W. Norton. Retrieved from https://editors.eol.org/eoearth/wiki/Species_richness

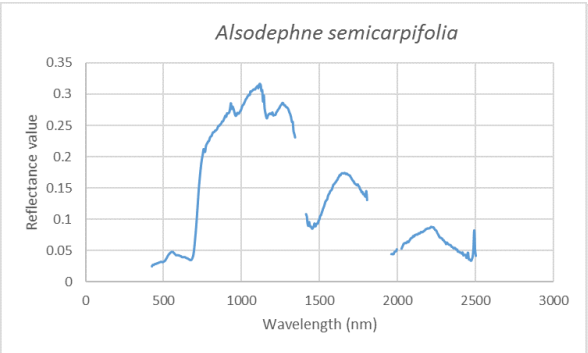
Xia, J., Bombrun, L., Berthoumieu, Y., Germain, C., & Du, P. (2017). Spectral–Spatial Rotation Forest for Hyperspectral Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(10), 4605–4613. https://doi.org/10.1109/JSTARS.2017.2720259

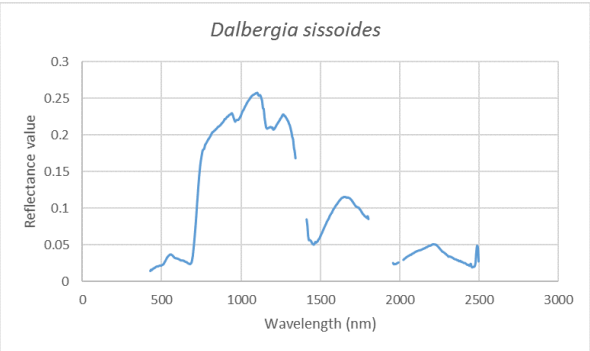
Youngentob, K. N., Roberts, D. A., Held, A. A., Dennison, P. E., Jia, X., & Lindenmayer, D. B. (2011). Mapping two Eucalyptus subgenera using multiple endmember spectral mixture analysis and continuum-removed imaging spectrometry data. *Remote Sensing of Environment*, *115*(5), 1115–1128. https://doi.org/10.1016/J.RSE.2010.12.012

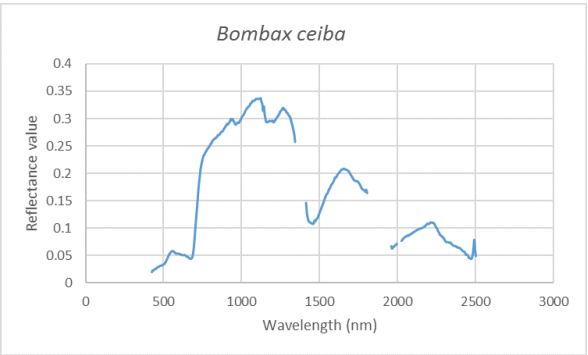
Zhang, L., & Suganthan, P. N. (2014). Random Forests with ensemble of feature spaces. *Pattern Recognition*, *47*(10), 3429–3437. https://doi.org/10.1016/J.PATCOG.2014.04.001

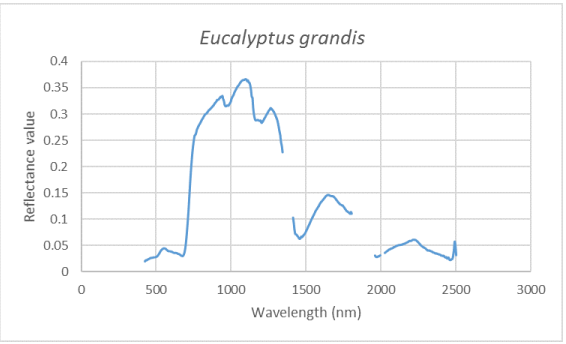
APPENDIX-A

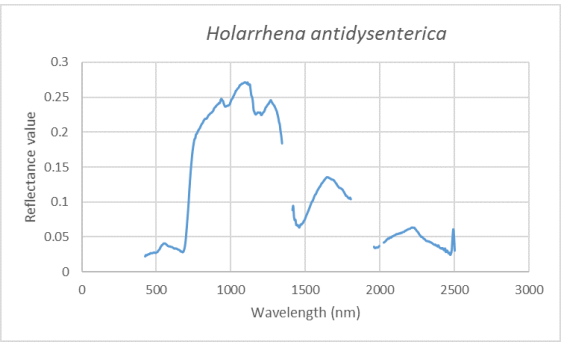
Following are the spectral signatures of all the tree species which were identified in Shimoga, Karnataka.

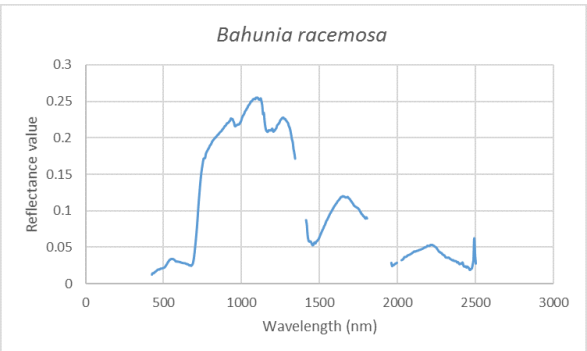


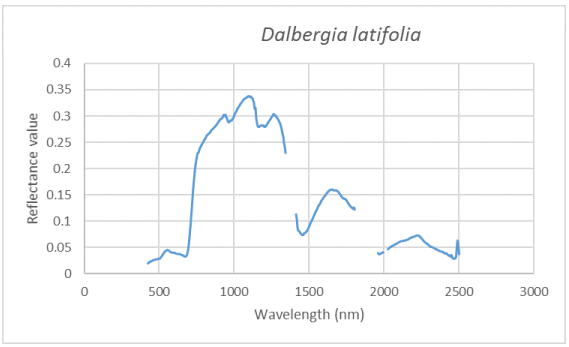


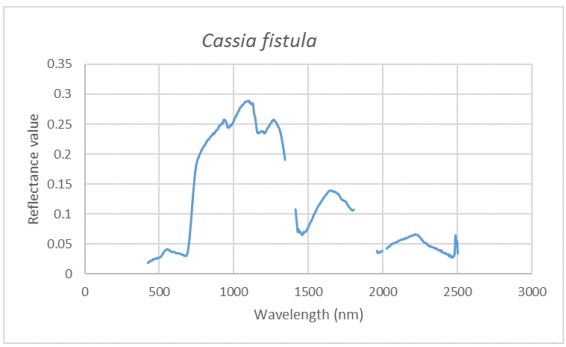


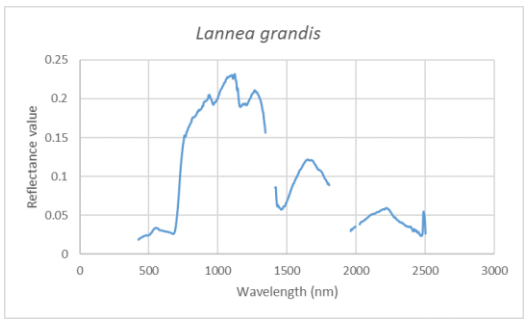


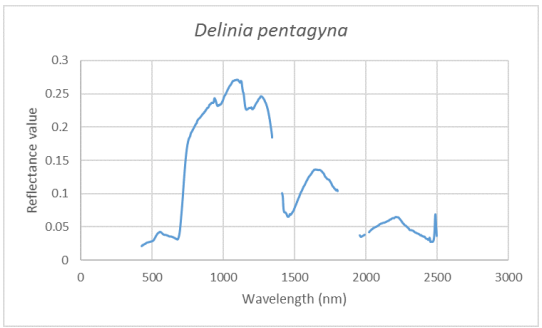


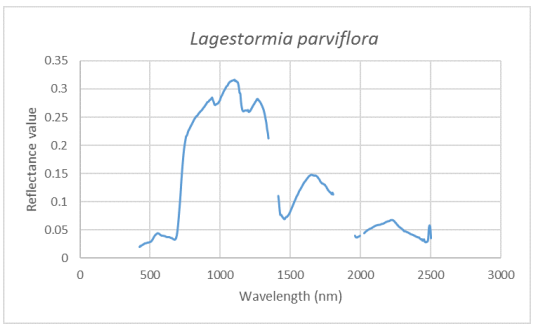


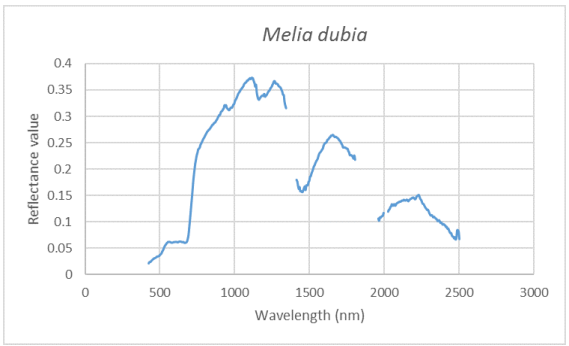

Dalbergia latifolia
Reflectance value
Wavelength (nm)

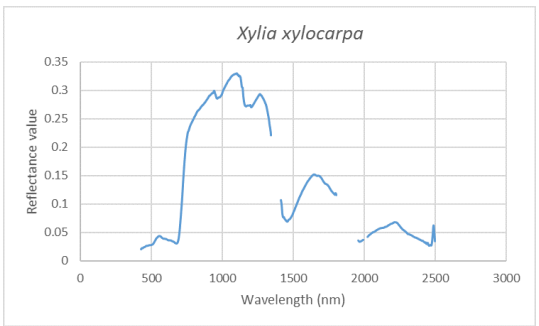

Cassia fistula
Reflectance value
Wavelength (nm)

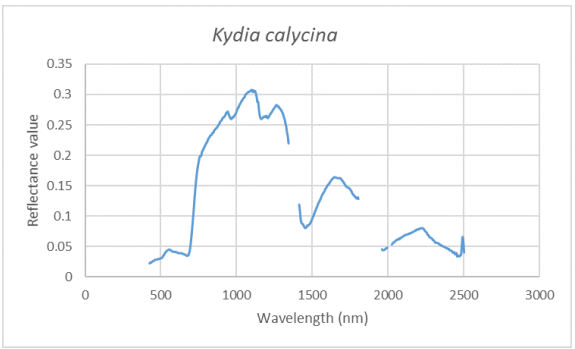

Lannea grandis
Reflectance value
Wavelength (nm)

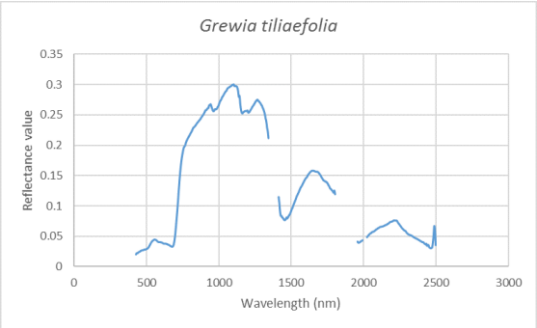

Delinia pentagyna
Reflectance value
Wavelength (nm)

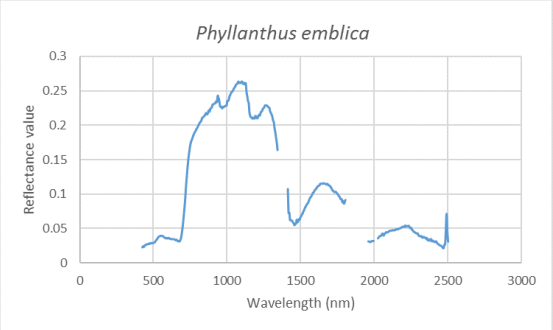

Lagestormia parviflora
Reflectance value
Wavelength (nm)

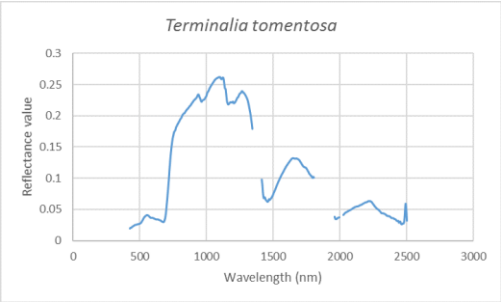

Melia dubia
Reflectance value
Wavelength (nm)

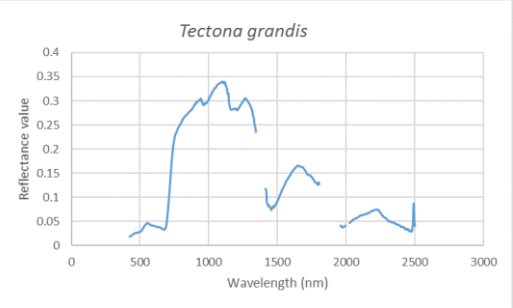

Xylia xylocarpa
Reflectance value
Wavelength (nm)

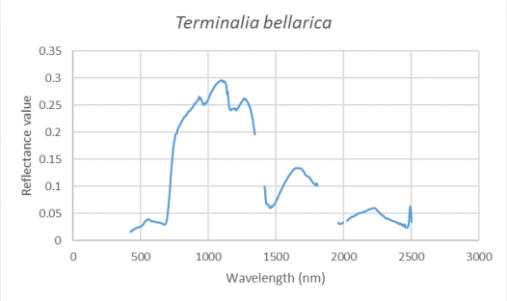

Kydia calycina
Reflectance value
Wavelength (nm)

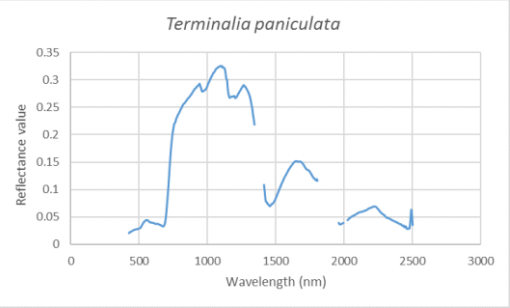

Grewia tiliaefolia
Reflectance value
Wavelength (nm)


Phyllanthus emblica
Reflectance value
Wavelength (nm)


Terminalia tomentosa
Reflectance value
Wavelength (nm)

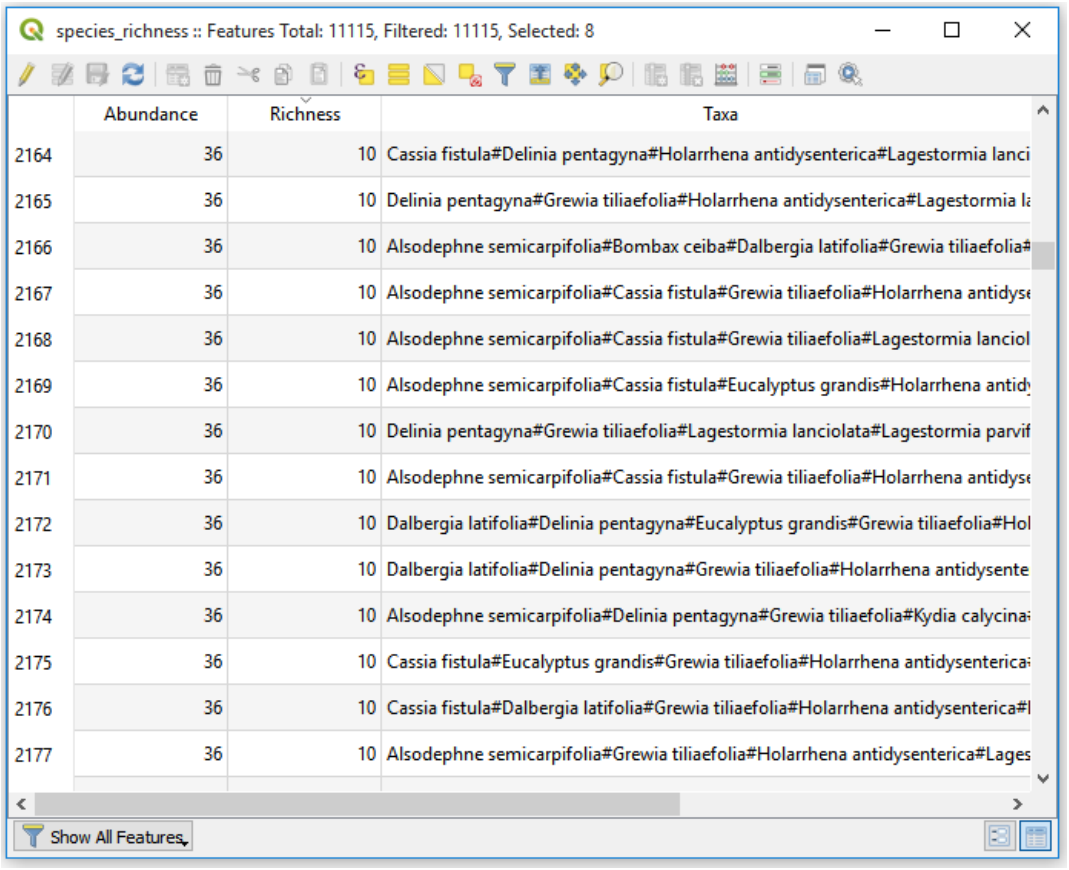

Tectona grandis
Reflectance value
Wavelength (nm)


Terminalia bellarica
Reflectance value
Wavelength (nm)


Terminalia paniculata
Reflectance value
Wavelength (nm)

APPENDIX-B

Following is the attribute table which was obtained using TomBio tool in QGIS provided with the information of species richness.



	Abundance	Richness	Taxa
2164	36	10	Cassia fistula#Delinia pentagyna#Holarrhena antidysenterica#Lagestormia lanci
2165	36	10	Delinia pentagyna#Grewia tiliaefolia#Holarrhena antidysenterica#Lagestormia la
2166	36	10	Alsodephne semicarpifolia#Bombax ceiba#Dalbergia latifolia#Grewia tiliaefolia#
2167	36	10	Alsodephne semicarpifolia#Cassia fistula#Grewia tiliaefolia#Holarrhena antidyse
2168	36	10	Alsodephne semicarpifolia#Cassia fistula#Grewia tiliaefolia#Lagestormia lanciol
2169	36	10	Alsodephne semicarpifolia#Cassia fistula#Eucalyptus grandis#Holarrhena antidy
2170	36	10	Delinia pentagyna#Grewia tiliaefolia#Lagestormia lanciolata#Lagestormia parvif
2171	36	10	Alsodephne semicarpifolia#Cassia fistula#Grewia tiliaefolia#Holarrhena antidyse
2172	36	10	Dalbergia latifolia#Delinia pentagyna#Eucalyptus grandis#Grewia tiliaefolia#Hol
2173	36	10	Dalbergia latifolia#Delinia pentagyna#Grewia tiliaefolia#Holarrhena antidysente
2174	36	10	Alsodephne semicarpifolia#Delinia pentagyna#Grewia tiliaefolia#Kydia calycina#
2175	36	10	Cassia fistula#Eucalyptus grandis#Grewia tiliaefolia#Holarrhena antidysenterica#
2176	36	10	Cassia fistula#Dalbergia latifolia#Grewia tiliaefolia#Holarrhena antidysenterica#I
2177	36	10	Alsodephne semicarpifolia#Grewia tiliaefolia#Holarrhena antidysenterica#Lages

APPENDIX-C

Following table gives the overall accuracy of all the three bands when all the bands were considered and when variable selection was done. After performing variable selection 20% of least important bands were removed and accuracy was calculated. Maximum accuracy was achieved after the first iteration in all three cases.

	All bands	**20 % bands removed**
SVM	41.21%	41.94%
RF	40.34%	41.88%
RoRF	52.76%	53.50%