CHAPTER 5: LEAF CHLOROPHYLL CONTENT RETRIEVAL FOR AVIRIS-NG IMAGERY USING DIFFERENT FEATURE SELECTION AND WAVELET ANALYSIS

5.1 Introduction:

LCC in a crop is a significant measure of its photosynthetic activity and nutrient level. As a result, it has been reported that they are the most significant organic pigment on the planet [26, 68]. A range of physiological stressors, leaf growth, and senescence cause changes in the overall concentration of chlorophyll (a & b). Nitrogen makes up a part of the chlorophyll molecule, thus the concentration of chlorophyll content is a proxy of plant nutritional status [69]. Thus, knowledge of spatial and temporal LCC is extremely valuable in studies of plant-environment interactions, agriculture, forestry, and environmental monitoring. Conventional approaches for obtaining this information are time-consuming and labour-intensive, relying on laboratory examination of foliar extracts, resulting in limited pigment information. The reflectance spectra of leaves show the spectral absorbance by pigments, which opens the possibility of using reflected radiation from field, airborne, and spaceborne instruments. These spectra obtained from various sensors can be used in a non-destructive method for quantifying pigments over a wide range of spatial scales and with greater temporal coverage than traditional sampling methods.

Hyperspectral reflectance data collected using field-based, airborne, and spaceborne equipment can measure vegetation chlorophyll concentrations at many spatial and temporal scales [70]. Most recent research on the detecting chlorophyll pigment using spectra analysis have concentrated on the investigation of spectral features to quantify the intensity and location of molecule absorption or reflection [71]. Many efforts have

been made to use the leaf surface reflectance in the extraction of LCC from remote sensing data. Various techniques for estimating leaf chlorophyll content are using parametric, non-parametric regression, physical, and hybrid models. One of the most extensively used parametric regression approaches is the use of spectral indices, which have been adjusted to employ reflectance in the blue wavelengths to lessen the impacts of surface reflectance on leaf pigment retrieval across various species [41, 72. Compared to regression approaches, physically-based models are made in the scientific community, offering accurate and robust predictions of pigment concentration across different species and spatial scales. Due to its broad validation and simplicity of inversion, the PROSPECT model is the most widely used leaf radiative transfer model {Baret, 1991 #46. Compared to regression approaches, physically-based models are made in the scientific community, offering accurate and robust predictions of pigment concentration across different species and spatial scales. Due to its broad validation and simplicity of inversion, the PROSPECT model is the most widely used leaf radiative transfer model {Baret, 1991 #46] Such models have been developed to comprehend radiation interactions with leaves and canopies and mimic reflectance spectra in forward mode. Iterative optimization and neural networks may be employed in inverse mode to predict the values of model inputs, such as pigment concentrations, that explains recorded spectra [73]. Many species- or site-specific approaches have resulted from a lack of explicit testing. Numerous studies that examined many spectrum techniques under various conditions found a lack of generality and usability [13], and there is little agreement on the ideal spectral approaches [74-76]. These investigations discovered that narrowband Simple Ratios (SR) and Normalized Difference Ratios (NDR) adjusted to include a waveband in the blue area for leaf surface reflectance were the best predictors of LCC. Although many

studies have been done using various pre-processing methods such as continuum removal and n^{th} order derivative, filters have been implemented to remove noise from the signals and enhance their characteristics that caused by factors like sample background and stray light [77, 78]. Summarily, hyperspectral remote sensing can meet the growing need for information on plant pigments at various spatial scales.

Wavelet Analysis (WA) was developed in numerous scientific domains, but interactions between them over the last decade have resulted in a varied spectrum of signal processing applications. New approaches in image compression, classification, and enhancement have identified WA's promise in image processing [79]. According to laboratory spectroscopic studies, WA significantly benefits over earlier spectrum techniques. Wavelets are mathematical functions that are used to represent data or other functions. Fourier analysis (FA) is an approximation via superposition of functions; however, data in WA is analysed at multiple scales [80]. Observing a spectrum via a big "window" detects massive characteristics, whereas observing it through a tiny "window" finds little features. For FA, sines and cosines are non-local and perform poorly when reflectance spectra have severe discontinuities. Since, WA can break down the signal into components localized in time and frequency domains, it can employ more appropriate functions to express local spectral aspects. In contrast, FA can only represent frequency information [81, 82]. Various spectroscopic studies suggest the usefulness of WA in removing background noise in reflectance to estimate chemical concentration accurately [83, 84]. It does so by preserving spectral features of components, which other filtering or smoothing algorithms have not found to be done efficiently. The WA is useful for differentiating signals with identical appearances but issued by systems with distinct dynamics [85]. Furthermore, WA has also proven its efficiency in classification problems; for example, Huang et

al.2001[86] discriminated soybean crops with grass weeds using a wavelet-based classification problem with an accuracy greater than 90%.

In the present study, several wavelets such as db, bior, and rbio of different levels were utilized to deconstruct and denoise the AVIRIS-NG spectra into multiple coefficients, as the use of WA to denoise vegetation spectra from an AVIRIS-NG imagery has yet to be studied for LCC retrieval. To further narrow down the numbers of relevant bands, a variety of feature selection techniques such as RFE, RRF, LASSO, and PLS were explored. Top three bands from each method were chosen to formulate SR and NDR kinds of VI. All possible combinations of bands were used to formulate the said VIs. Afterwards, in-situ LCC was trained and validated using these indices.

5.2 Methodology:

5.2.1 Study site:

The study site for the current work is situated in Anand district, Gujarat as discussed in Chapter 2, Section 2.3.1, and shown in Figure 2.8. The land cover is primarily for agricultural use, to grow crops such as wheat, tobacco, maize, and other pulses. AVIRIS-NG flight scene for the study area along with sampling points is shown in Figure 5.1.

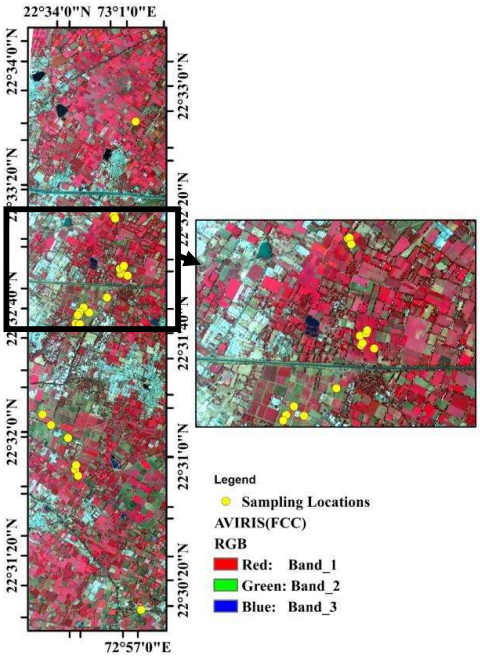

Figure 5.1. AVIRIS-NG coverage with a zoomed image of the campaign site

5.2.2 Datasets:

An extensive field sampling of the study site was conducted coinciding with the dates of flight AVIRIS-NG overpass for 2016, February 7. Garmin eTrex 10 GPS was used to locate the ground data points for various agricultural sites corresponding to various crop types over twenty-six locations on the campaign site. Among those crop types, nine were identified during the study area sampling. The level-2 surface reflectance for the AVIRIS NG was downloaded from VEDAS-SAC geoportal for February 2016. The presence of different gases and aerosols in the atmosphere makes the data noisy and needs to be taken care of using atmospheric correction approaches. The reflectance data acquired from AVIRIS NG was atmospherically corrected to generate L0 and L1 products (radiance to reflectance) [33, 87]. The LCC of the vegetation was recorded

using SPAD. The wavelet analysis was done in MatLab R2021a; the feature selection process was executed in Rstudio.

The graphical representation of the entire work is presented in Figure 5.2.

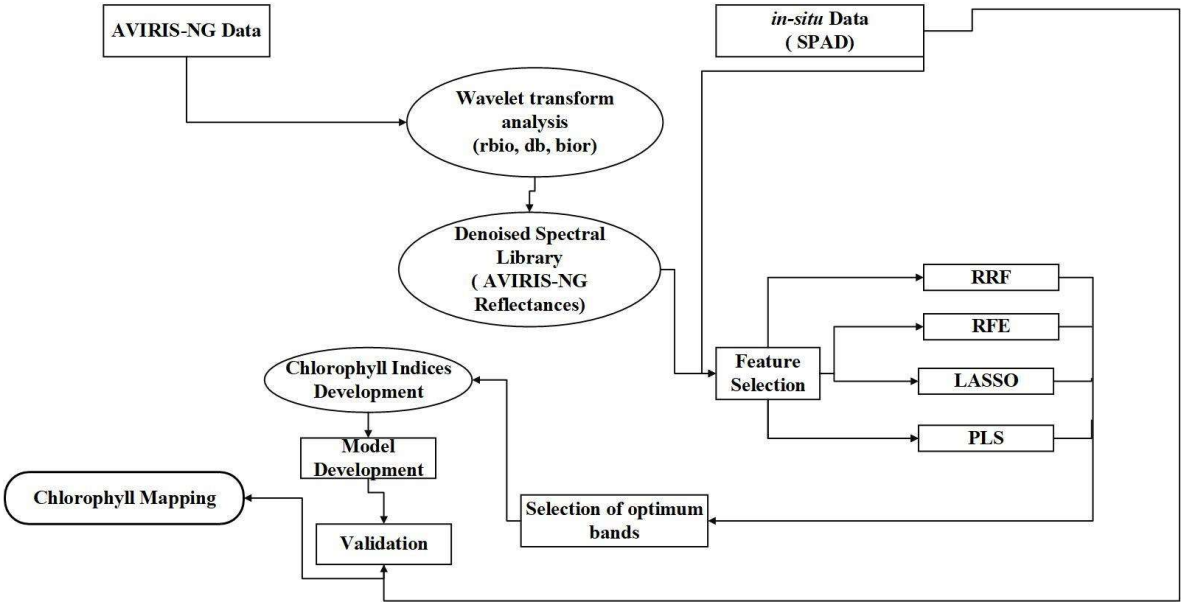

Figure 5.2. Flowchart of the methodology for LCC retrieval

5.2.3 Wavelet analysis (WA):

Different types of wavelet analysis were used over the spectrum to denoise the vegetation's spectra for the retrieval of LCC. Several criteria distinguish the types of wavelets, including if they have compact support, symmetry, regularity, and a scaling function, as well as the number of vanishing moments, and the orthogonality or biorthogonality of the subsequent analysis. In order to choose the most appropriate wavelet for an application, the features like relationship between the shape of the wavelet and the shape of the signal of interest are compared. As per Blackburn, 2007 [88], to avoid the problem of overfitting and avoid a large volume of processing, coefficient details from level 3 and 5 was chosen to run over the vegetation spectra.

Accordingly, for this study, level 3 was chosen for the analysis because wavelets for level 4 and 5 the approximation performance showed a poor relation based on RMSE. Wavelets like Haar, Coiflets and Symlets were giving poor approximation for the vegetation signal, whereas wavelets like, db, bior, and rbio gave fair results in comparison to another wavelet family. Furthermore, after running the wavelets analysis with different kinds of wavelets, bior3.7 gave the best approximation for vegetation spectra, followed by, rbio3.7, and db7.

5.2.4 Feature selection methods (FSM):

Feature selection (FS) is a machine learning strategy for selecting a subset of relevant information, specifically variables, for model generation. The feature selection approach seeks to remove redundant or unnecessary characteristics or features highly linked in the data with little information loss. FS analysis was done between the dataset's approximated coefficients and LCC values to identify the top three bands most relevant to LCC based on Variable Importance score (VIS). VIS is determined by sum of the decrease in error when split by a variable. Different FSM were chosen to find out that which would give the best model for LCC estimation, these FSMs are:

5.2.4.1 Recursive Feature Elimination (RFE)

RFE is most popular FS algorithm because it is simple to use and also effective to select the most optimum features from a training dataset which will be more relevant in predicting the target variable. It is a wrapper-style algorithm, which specifies that during the process, different types of machine learning algorithm employed in the method's core and further it is wrapped with RFE for feature selection. RFE or filter-based feature selection algorithm assigns scores to each feature and choose the optimum feature with the highest (or lowest) value. Since, RFE wrapper-style feature

selection method has additionally filter-based feature selection method internally. In order to do this, the specified machine learning method employed in the model's core is fitted, the features are ranked by relevance, and the least important features are discarded.

RFE removes *n* features from a model by iteratively fitting it, and at each iteration, deleting the weakest features, as determined by the coefficient or feature importance's attribute of the fitted model [89].

5.2.4.2 Regularised Random Forest (RRF)

New and most advanced RRF was introduced for FS using only one ensemble. It evaluates features on a subset of the training data at each tree node. It derives an upper bound on the number of different gain values in a node and show that numerous features at a node with a small number of instances and a large number of features can share the same information gain. As a result, in a node with a small number of instances, RRF is likely to choose a feature that is not highly significant. The gain in RRF for each feature X_k is denoted as:

$$G_{RRF}(X_k, v) = G(X_k, v) \quad if \quad k \in F\} \tag{5.1}$$

$$G_{RRF}(X_k, v) = \lambda G(X_k, v) \quad if \quad k \notin F\} \tag{5.2}$$

F used for the chosen feature set that also useful to divide the instances of the preceding nodes, and $\lambda \in [0, 1]$ represents a penalty factor for the non-chosen features of the preceding nodes. RRF is able to choose non-redundant features because features with zero gain values are excluded from the selected features [90].

5.2.4.3 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO was developed in 1996 [91], for the estimation of parameters and feature selection in regression study. The method is a specific case of the penalized least squares regression with L1-penalty function.

LASSO estimate can be defined by

$$\hat{\beta}^{lasso} = \arg min \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (5.3)$$

Which can be written as

$$\hat{\beta}^{lasso} = \arg min \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2, \quad (5.4)$$

$$subject\ to \sum_{j=1}^{p} |\beta_j| \leq t,$$

the right-hand side of the equation can be broken down as (Residual sum of squares) + λ*(sum of the absolute value of the magnitude of coefficients), where λ is a parameter and denotes shrinkage. Shrinkage is when data values are minimized towards a central point as a mean, $\lambda = 0$, which means all features are considered, and it becomes linear regression, $\lambda = \infty$, which means no feature is considered.

5.2.4.4 Partial Least Square (PLS)

PLSR is a statistical approach that was established in 1966 [92]. It associates data matrices using a linear multivariate model, decreasing collinearity and noise in a dataset [93]. Instead of depending on the original data, this two-step strategy lowers predictors of a dataset to a smaller number of uncorrelated components before running least squares regression on them. The principal component and multilinear regression properties are generalized and combined in PLSR. The tool comes in handy when there are several variables and a lot of collinearity between them. This regression generates the following linear model:

$$Y = X\beta+\varepsilon \quad (5.5)$$

where Y is a mean-centered vector of a dependent variable, X is a mean-centered matrix of independent variables, β and ε are regression coefficient and residual matrices, respectively. The projection variable relevance assesses the contribution of independent variables to dependent variables.

5.3 LCC retrieval

For the retrieval of LCC using the best bands and the generated indices, linear regression model was developed based on the relationships between two variables by fitting a linear equation to observed SPAD datasets and indices. In this case, one variable was selected as an explanatory variable (x_i), while the other chosen as dependent variable (y_i) [94]and can be represented as:

$$y_i = x_i b + e_i, \quad (5.6)$$

where $i = 1, \ldots, n,;$ b = vector of unknown parameters with e_i as stochastic disturbances. Over here, the Generalized Linear Model (GLM) has been used to represent the flexible generalization of the linear regression equation by allowing the response variables having error distribution instead of having the normal distribution. The GLM model used in this study comprises of stochastic, systematic components and afterwards a linkage function was used to link random and systematic components [95].

For assessing the performances of the model outputs, the Pearson's correlation coefficients (R) was used between the best indices and LCC.

5.4 RESULTS AND DISCUSSION:

5.4.1 Wavelet analysis

The presented work estimated LCC values based on the vegetation index model derived for AVIRIS-NG imagery. Before developing the LCC model, the vegetation spectra were scrutinized for removing the noise using WA to give best retrieval model possible. Different families of wavelets were used for carrying out WA. Amongst all levels, level 3 wavelet types were used to avoid large amount of intensive processing, based on the study by Blackburn [88]. The wavelet toolbox of MatLab offers eight types of wavelets such as bior, db, rbio, haar, sym, coif, dmey, and fk. All of these wavelets for level 3 have been applied on the raw vegetation spectra. We implemented WA of each order on the spectra and to determine the best wavelet to be used, calculated RMSE between approximated signal and original signal was computed. Based on RMSE values, bior3.7, db7, and rbio3.7 topped the list and hence bior3.7 wavelet was chosen for further analysis. RMSE values for each wavelet can be seen in Table 5.1. Figure 5.3 to 5.5 shows the decomposition at three levels and approximation of one of the signals for different wavelets

Table 5.1: RMSE values of each wavelet type for level 3

Wavelet	**bior3.7**	**db7**	**rbio3.7**	**haar**	**sym8**	**coif3**	**dmey**	**fk4**
RMSE	0.011	1.257	1.011	3.304	2.102	3.209	2.264	2.208

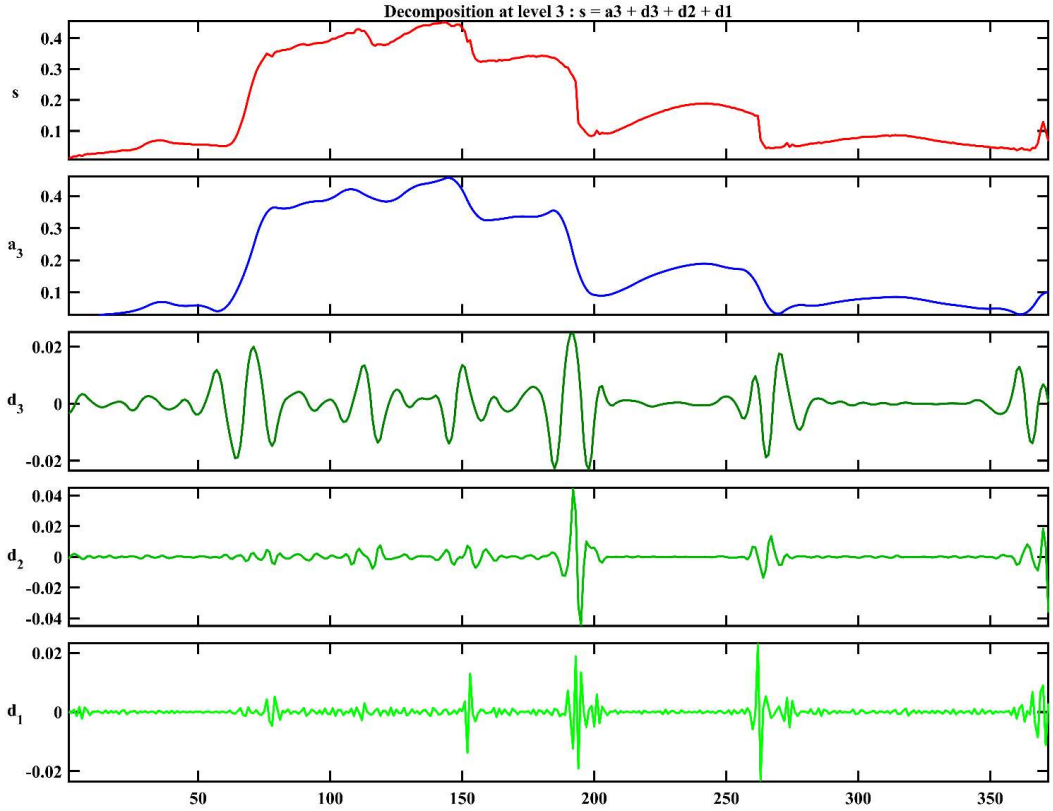

Figure 5.3. WA of a vegetation spectra using bior3.7 at level 3

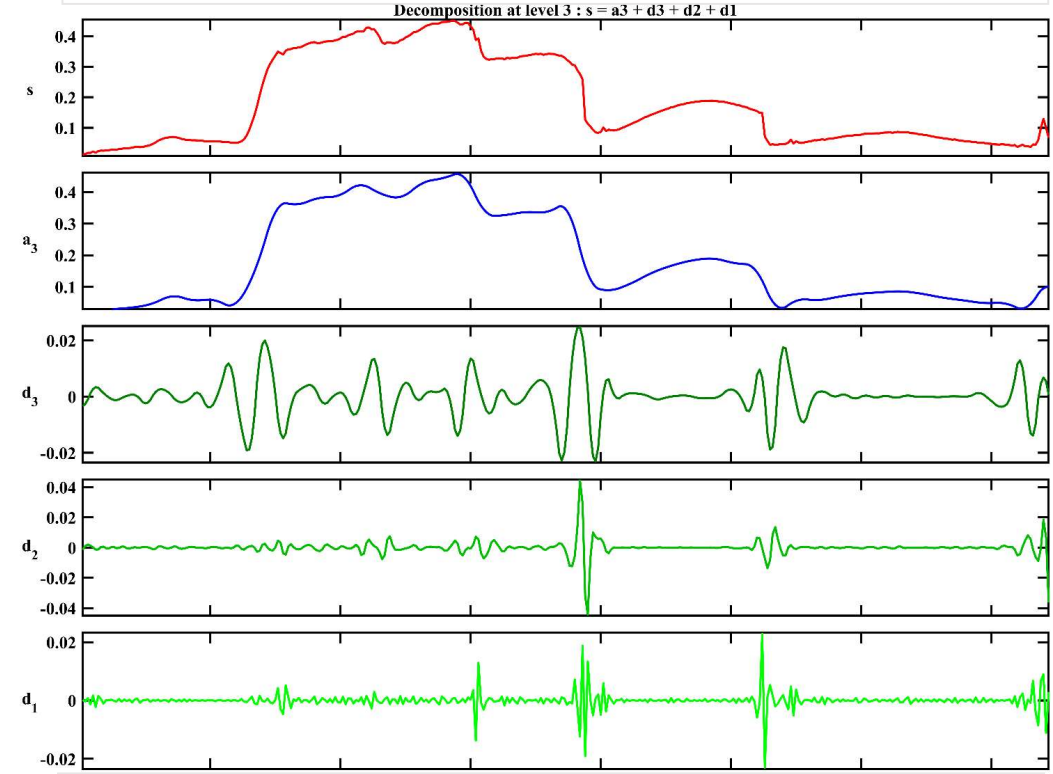

Figure 5.4. WA of a vegetation spectra using rbio3.7 at level 3

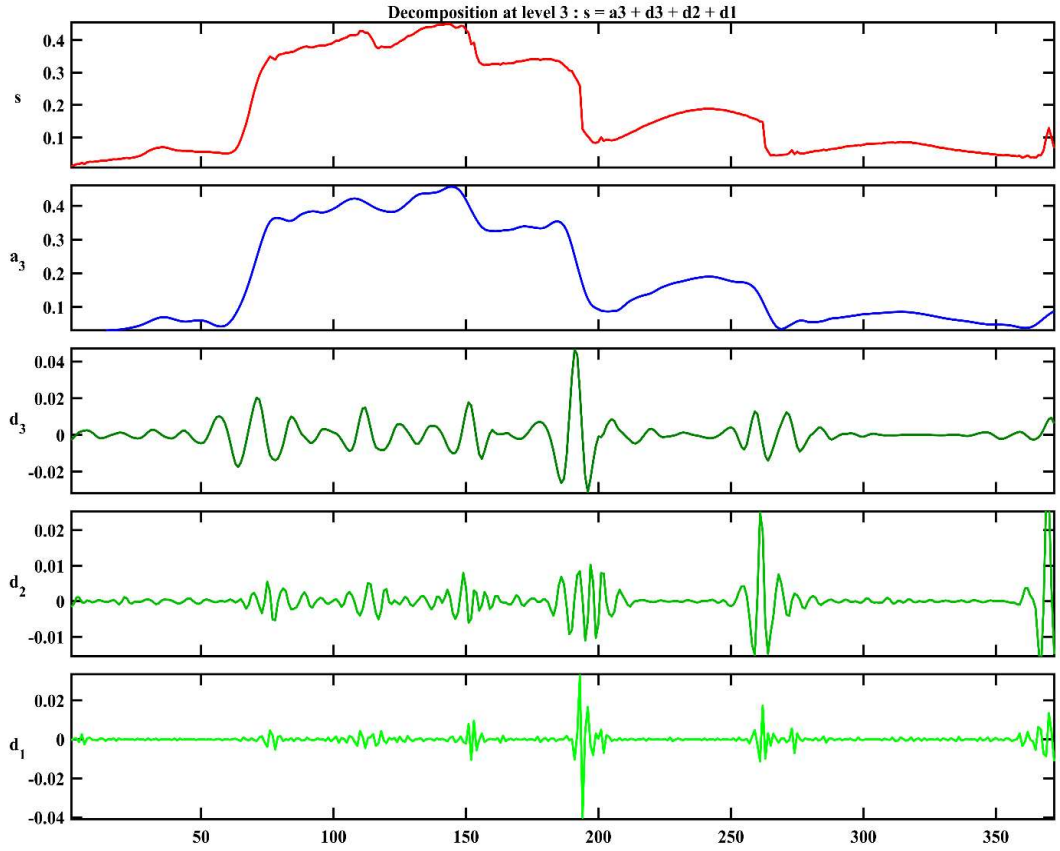

Figure 5.5. WA of a vegetation spectra using db7 at level 3

In the Figures 5.3-5.5, 's' represents signal, a_3 is approximation coefficients at level 3, and d_1, d_2, d_3 are decomposition coefficients at 3 levels. The wavelet decomposes the signal into its constituent signals, the noisy signals have very low value which are removed, and rest of the signals are combined back to give denoised signal. The best approximation on the basis of Root Mean Square Error (RMSE) was bior3,.7, which was found to be 0.011, and, hence, bior3.7, level 3 was used for further analysis. The denoised signal was further given to different FSMs to identify important bands for the estimation of LCC.

5.4.2 Selection of optimum bands:

The primary motivation for using FS algorithm is to identify and remove as much irrelevant and redundant information as possible, which is a significant challenge with hyperspectral data due to its complex and high dimensional distribution. The FS algorithms not only aid in achieving the desired efficiency, but they are also time efficient when it comes to find out the best bands among hundreds of bands. Figure 5.6(a-d) show the graph of comparison of VIS for bands using different FS algorithms.

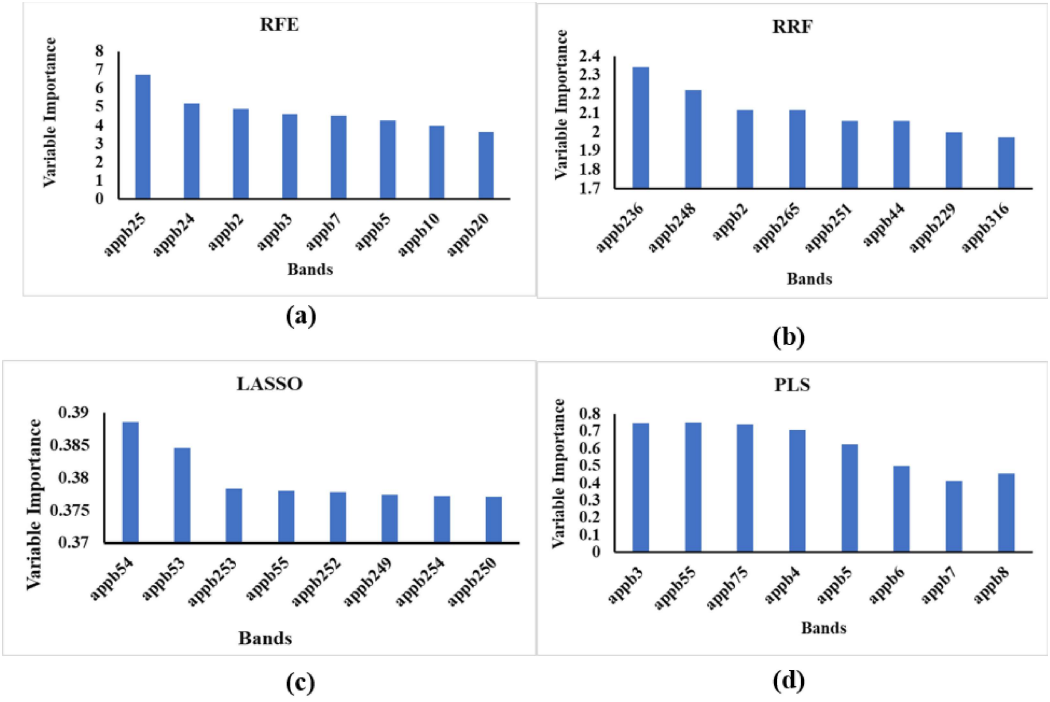

Figure 5.6 (a-d). Variable importance from different FS algorithms

In Figures 5.6(a-d), 'appbx' denotes approximated band no. x, identified by FS algorithms, plotted against VIS. For RFE, appb25, appb24, and appb2 i.e 497nm, 492nm, and 381.8nm bands respectively, are the most significant bands chosen, with VIS of 6.73, 5.20, and 4.88. Using RRF, appb236, appb248, and appb2 (1554nm, 1614nm, and 381.8nm bands, respectively) are the most important bands, with VIS of

2.34, 2.22, and 2.12, respectively. Although LASSO regression gave the worst results with negative correlations with LCC values, bands 54, 53, and 253 (642.3nm, 637.3nm, and 1639nm, respectively) were important according to this method. The best model output was given by the indices generated with PLS methods, according to which approximated bands 3 (386.8nm), 55 (647.3nm), and 75 (747.5nm) were the most significant bands with VIS of 0.75, 0.74, and 0.73, respectively. These bands fall in blue, red and NIR region of the electromagnetic spectrum.

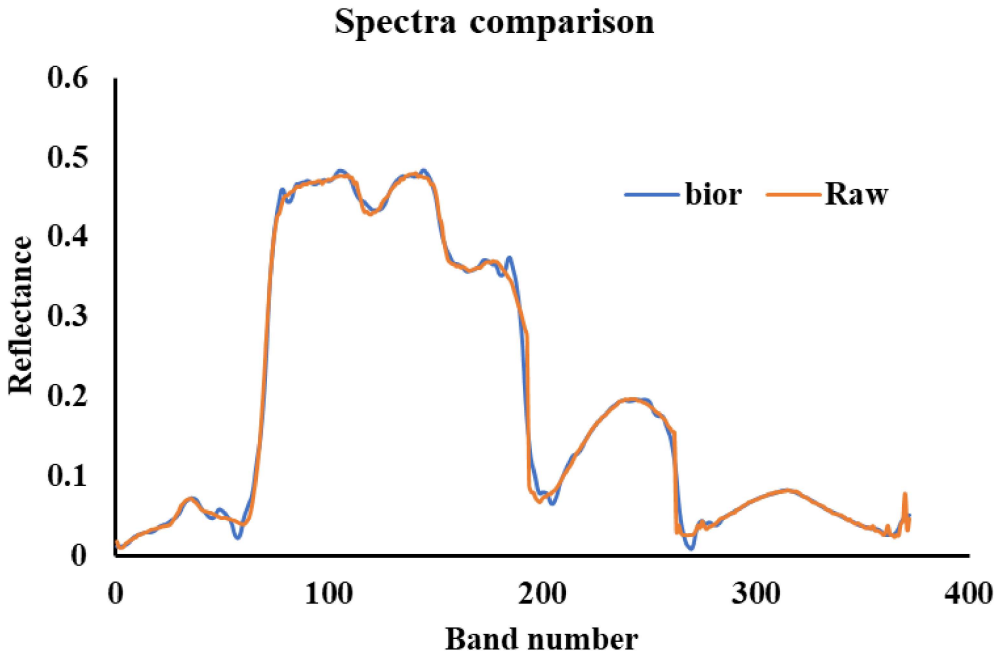

Figure 5.7. Comparison between raw spectra and denoised spectra using bior3.7

Figure 5.7 shows the comparison between a vegetation spectrum from the AVIRIS-NG image and denoised spectra for the same point using bior3.7 wavelet. It is evident that the denoised spectra is smoother than the original spectra.

5.4.3 LCC mapping:

After selecting bands important to LCC, the top three bands were selected for forming two types of VIs, namely, Simple Ratios (SR) and Normalized Difference Ratios (NDR), for each possible band combination. The correlation between LCC corresponding to each indices value is shown in Table 5.2. The VIs giving the best correlation values w.r.t. LCC were selected to train the model. Among all, 70% of the data was used to train the linear regression model, and the rest 30% of the data was used to validate it. The performances of the model were evaluated in terms of correlation obtained between the selected band ratios and LCC. The highest performing model was then used for generation of large-scale chlorophyll inventory of the area as shown in Figure 5.9.

Table 5.2. Correlation between each calculated VIs from all the methods w.r.t LCC values

RRF (Regularized Random Forest)		**RFE (Recursive Feature Elimination)**	
Simple Ratio (SR(X, Y))	Correlation (R)	**Simple Ratio (SR(X, Y))**	**Correlation (R)**
SR (2,236)	-0.600	SR (24,25)	-0.500
SR (2,248)	-0.602	SR (24,2)	0.783
SR (236,248)	-0.250	**SR (25,2)**	**0.784**
SR (236,2)	0.713	SR (25,24)	0.501
SR (248,2)	0.718	SR (2,24)	-0.757
SR (248,236)	0.248	SR (2,25)	-0.755
Normalized Difference	Correlation (R)	**Normalized Difference**	**Correlation (R)**

Ratio (NDR(X,Y))		**Ratio (NDR(X,Y))**	
NDR (2,236)	-0.613	NDR (24,25)	-0.500
NDR (2,248)	-0.616	**NDR (24,2)**	**0.779**
NDR (236,248)	-0.249	NDR (25,2)	0.777
NDR (236,2)	0.613	NDR (25,24)	0.500
NDR (248,2)	0.616	NDR (2,24)	-0.779
NDR (248,236)	0.249	NDR (2,25)	-0.777

LASSO		**PLS (Partial Least Square)**	
Simple Ratio (SR(X, Y))	Correlation (R)	**Simple Ratio (SR(X, Y))**	**Correlation (R)**
SR (54,53)	-0.202	SR (75,3)	-0.809
SR (54,253)	-0.143	SR (55,75)	-0.807
SR (53,253)	-0.125	SR (3,75)	-0.795
SR (53,54)	0.199	SR (3,55)	0.808
SR (253,54)	0.125	SR (75,55)	0.805
SR (253,53)	0.110	SR (75,3)	0.794
Normalized Difference Ratio (NDR(X, Y))	Correlation (R)	**Normalized Difference Ratio (NDR(X, Y))**	**Correlation (R)**
NDR (54,53)	-0.201	NDR (55,3)	-0.809
NDR (54,253)	-0.139	NDR (55,75)	-0.806
NDR (53,253)	-0.122	NDR (3,75)	-0.795

NDR (53,54)	0.201	**NDR (3,55)**	**0.809**
NDR (253,54)	0.139	NDR (75,55)	0.806
NDR (253,53)	0.122	NDR (75,3)	0.795

For the top three bands identified by four FSM, SR, and NDR for each band combinations were calculated, 6 SR indices and 6 NDR indices were calculated for three bands, (in total 48 indices were calculated by combining each FS method). In Table 5.2, SR (X,Y) and NDR (X,Y) are Simple Ratio and Normalized Difference Ratio VIs between approximated band numbers X and Y. Correlations for all the indices with LCC values were checked for each method, and highly correlated indices were selected to form the LCC retrieval method (highlighted in Table 5.2 in blue).Highly correlated VI from RRF was SR between band 248, and 2 with R value of 0.718; for RFE it is SR between bands 25 and 2, followed by NDR between 24 and 2. It is observed that LASSO performed poorly among all, whereas for PLS the highly correlated VI w.r.t. LCC is NDR formed between bands 3 and 55. The linear regression model for each highly correlated model is shown in Table 5.3, and the corresponding Taylor plot (Figure 5.8) is shown to identify the best model. It is clear from Table 5.3 and Figure 5.8, that PLS method gives the best model using NDR (3,55) bands 3 and 55, normalized ratio between band 386.8nm and 647.3nm for the LCC retrieval with correlation (r) value of 0.94.

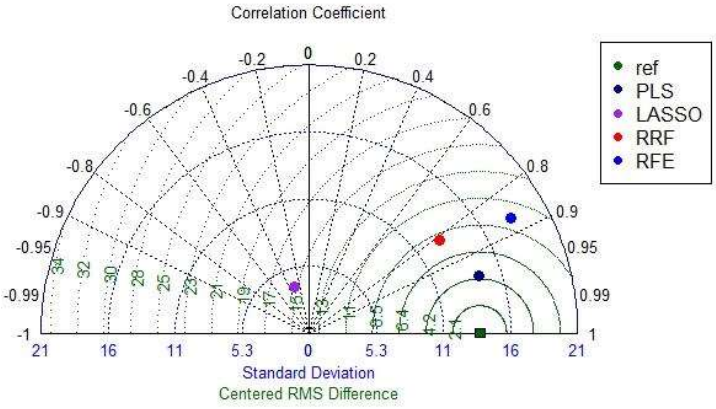

Figure 5.8. Taylor plot comparing models

Table 5.3. LCC retrieval models

FSM	Models	Validation (R)
RRF	Y = 1.7644*SR (53,54) + 22.142	0.81
RFE	Y = 10.695*SR (25,2) + 10.73	0.87
LASSO	Y= 77.631*NDR (53,54) + 8.857	-0.31
PLS	Y = 16313* NDR (3,55) + 24.4	0.94

The PLS method-derived model was chosen to make chlorophyll map for the AVIRIS-NG imagery (Figure 5.9).

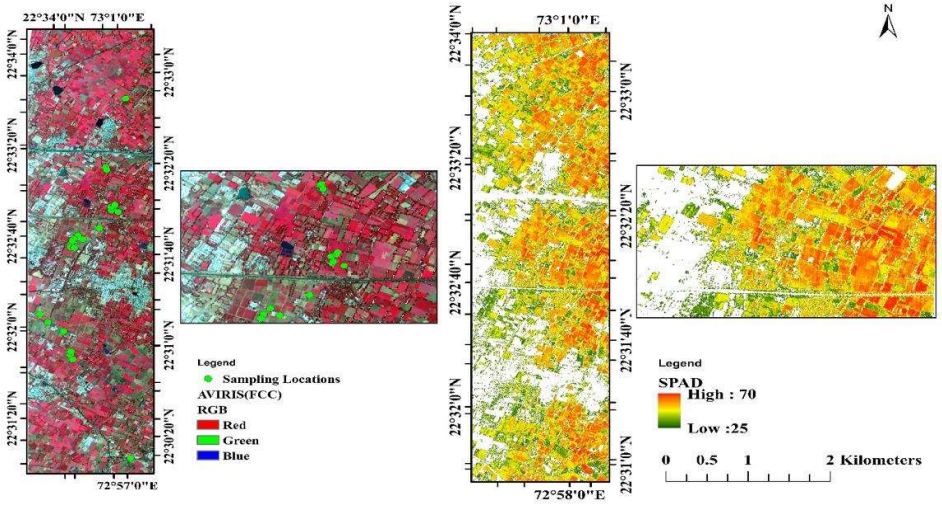

Figure 5.9. Chlorophyll map for the entire AVIRIS-NG imagery.

In Figure 5.9, it is evident from a small area from the study site has been zoomed out that shows all the non-agricultural areas (urban, road, soil) have been properly masked. Later, from the chlorophyll map, rest 30% data points were validated from the in-situ LCC values shown in Figure 5.10.

Using PLS feature selection algorithm, bands 3, 55, 75 were chosen, (bands with wavelength at 386nm, 647nm, and 747nm) and NDR for bands 3 and 55 gave the highest correlation for LCC. Subsequently, after using 70% of the data to train the linear model and rest 30% of the data to validate the model, the PLS FS algorithm performed better and topped the chart with R = 0.948, RMSE = 5.464, and bias = 3.305 values, It could be explained by the fact for chlorophyll content visible and NIR bands are very much relevant.

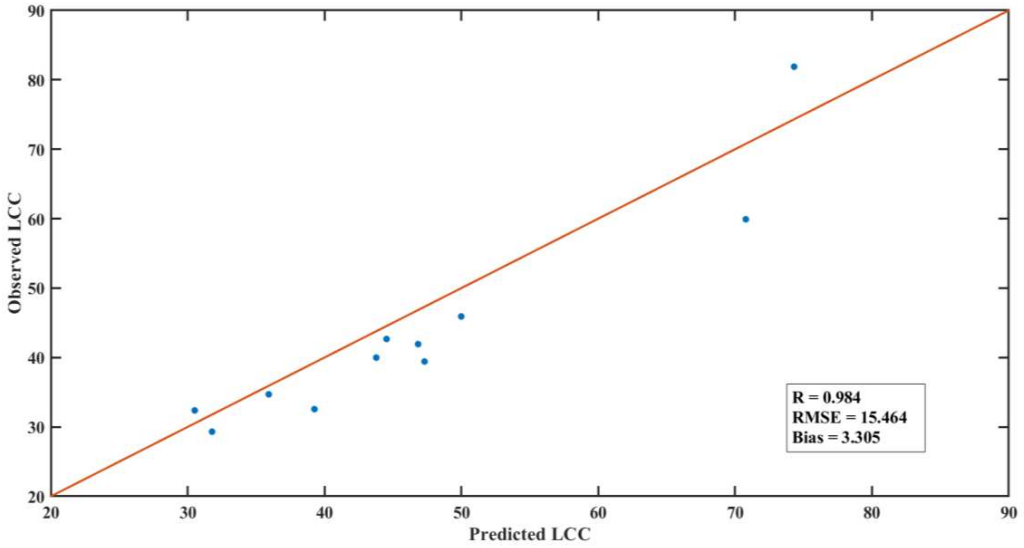

Figure 5.10. Comparison of estimated and measured LCC

This result is also supported by many previous studies that the bands in visible and NIR region have sensitivity towards LCC. Many vegetation indices such as Chlorophyll Absorption Ratio Index (CARI), Modified Chlorophyll Absorption Ratio Index (MCARI), MCARI/OSAVI, MCARI2, etc. are formed by bands in visible, and NIR bands, [75, 96-99]. From Table 5.2, it is clear that for RFE, highly correlated index is SR (25,2) while the next in same category goes to SR (24,2). Although there is difference of only 5nm between both bands, it is clear that denoising them using WA works out well even with very small difference between bands. Similar explanation goes for RFE: NDR (24,2) and RFE:NDR(25,2).

According to Blackburn, 2007, in wavelet analysis DWT is much better for Noise removal of the signals, and that is why different DWT were chosen for the analysis. Amongst all wavelets Biorthogonal wavelet was found best due to its similar shape with the signal used. It is also found that level 3 wavelets are found to be the best, which agrees with our study.

5.5 CONCLUSIONS

The present work attempt to show the usefulness of wavelet analysis and different feature selection methods to generate a model for the retrieval of LCC with AVIRIS-NG image. For denoising of spectra, DWT methods were used such as Daubechies, biorthogonal, reverse biorthogonal of level 3. Using bior3.7, level 3 we got the best approximated bands. These bands were further scrutinized to identify most significant ones using four FSM, namely, RFE, RRF, LASSO, and PLS. Top three bands were chosen from each method, to make two types of VIs for each possible band combination, namely SR and NDR. Amongst 48 developed indices, top 4 indices (one index from each method) were used to form a linear model to estimate LCC. Further, 70 % of the dataset were used to train the model and rest of the dataset were used to validate model. From the study and comparison with previous study, it is evident that denoising the spectra decreases marginal error in the estimation of LCC by clearing out insignificant noise present in the signal. Model generated by normalized ratio between band 3 i.e. blue region and band 55 i.e. NIR region using PLS method proved to be the best method with R = 0.948, RMSE = 5.464, and bias = 3.305 values. So, in overall, we can say that wavelet analysis combined with FSM has a very good potential for the estimation of LCC in vegetation.