THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

DEVELOPING A PREDICTIVE MODELLING APPROACH FOR REMOTE MEASUREMENT OF LEAF PHOTOSYNTHETIC FUNCTION

By
Hrishikesh Patel

School of Information Technology and Electrical Engineering,
The University of Queensland.

Final Report

Submitted for Master of Data Science Capstone 2

7th June 2021

Academic Supervisor : Prof. Diane Donovan

Executive Summary / Abstract

Increasing global warming is a significant issue today. However, reducing carbon emissions can alleviate it. Though plants play an essential role in sinking carbon on the land (Friedlingstein et al., 2019), the measurement of carbon sinks is highly uncertain. Hence, we need an accurate method to track carbon sinks. By monitoring plant photosynthesis, we can track carbon sinks provided by plants. This report outlines the data science approach used to develop a machine learning model to predict photosynthesis. The study involved making a predictive model to accurately predict photosynthesis function from Sun-Induced Fluorescence (SIF) data. Furthermore, the study also identified a small number of optimal wavelength bands from full wavelengths of the SIF data to predict photosynthetic capacity. Mainly, I trained the random forest regressor model on the training data and selected the best-performing model using cross-validation. I evaluated its performance on test data. Additionally, I compared the model with a univariate linear regression model. Furthermore, I recursively removed the least important variables (wavelength bands) from the model using recursive feature elimination to identify optimal bands. The study has found that the model could explain 86% of the variance in the target variable (YPSII) using the input variables. The model also performed better than the univariate linear regression model by exceeding its accuracy by 30%. Furthermore, I also identified optimal 40 wavelengths from the total 184 wavelengths. The model trained using these 40 wavelengths as input variables also found to be 86% accurate. These findings imply that SIF data has shown promising potential to predict the photosynthesis process accurately. Since a small dataset limited the modeling, I recommended using a larger dataset to obtain robust models. Furthermore, I also recommend trying various models such as Neural network, Xgboost, etc., to improve predictive accuracy (Ray, 2017). Although physical-based models are available to predict photosynthesis, the highly accurate machine learning models will improve confidence in the predictions. The successful modeling will be a positive step towards reducing carbon uncertainty and hence alleviating global warming.

Table of Contents

1. Introduction

In the upcoming decades, two significant problems are expected to surface. First, the amount of farmland will decrease, and food demand will increase due to the rising population. Since more than eighty percent of food calories are plant-based (Ritchie, 2017), tracking plant health has become increasingly necessary to maintain plant productivity. Plant health can be tracked by monitoring a photosynthesis process in plants (Garner, 2013).

The second major problem is the expected rise in global temperature due to the emission of greenhouse gases such as Carbon Dioxide (CO_2). Plants are a major sink of carbon dioxide on the land (Friedlingstein et al., 2019). Primarily, plants use CO_2 in the photosynthesis process. By monitoring photosynthesis, we can estimate the exchange of atmospheric CO_2.

Hence, active monitoring of photosynthesis in plants is necessary to track plant health and carbon exchange. Furthermore, due to advancements in satellite technologies, spectral measurements of plant properties have now become easier than a couple of decades ago (Guanter et al., 2015). There have been significant research in estimating photosynthesis from spectral data. However, very few researchers have used sun-induced fluorescence (SIF) to track photosynthesis. The project aims to link SIF and photosynthesis using machine learning techniques.

1.1 Background

Plants use absorbed sunlight to perform the photosynthesis process. In the process, plants react with water and carbon dioxide in the presence of sunlight to generate oxygen and food. However, when excess sunlight, plants use other mechanisms in addition to photosynthesis to handle the light (Garner, 2013). Figure-1 illustrates these mechanisms. The plant can emit the extra sunlight as heat or as fluorescence. By tracking the fluorescence, we can indirectly measure photosynthesis (Garner, 2013).

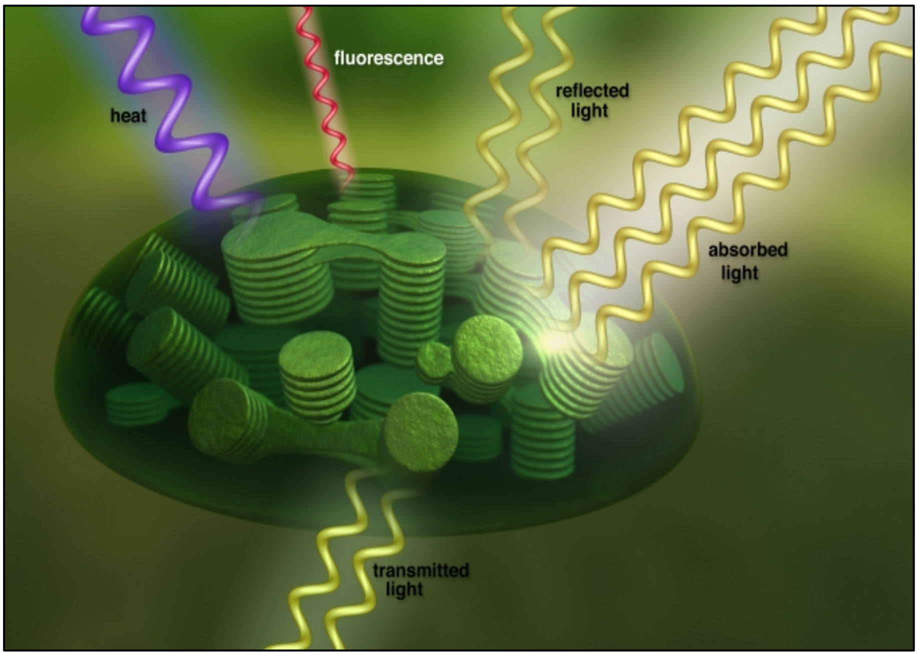


Figure 1 Mechanisms in plants to handle excess sunlight (Garner, 2013)

The fluorescence or sun-induced fluorescence is a glow of light that is not visible through human eyes. The higher SIF indicates active photosynthesis and a healthy plant, whereas low SIF can mean the plant is stressed (Garner, 2013). The SIF originates from two photosynthetic machineries, namely photosystem-1 (PS-1) and photosystem-2 (PS-2), as shown in figure-2. The photosynthetic yield of PS-2 is called YPSII, which I used in our data as a photosynthetic capacity. Furthermore, the SIF is highly sensitive to surrounding light conditions, as evident from figure-3. It can be observed that SIF values are higher in dark or low light conditions, where the plant is not stressed, and photosynthesis is efficient and vice versa for very bright conditions (Mohammed et al., 2019). This observation implies that SIF can serve as a better predictor for photosynthesis.

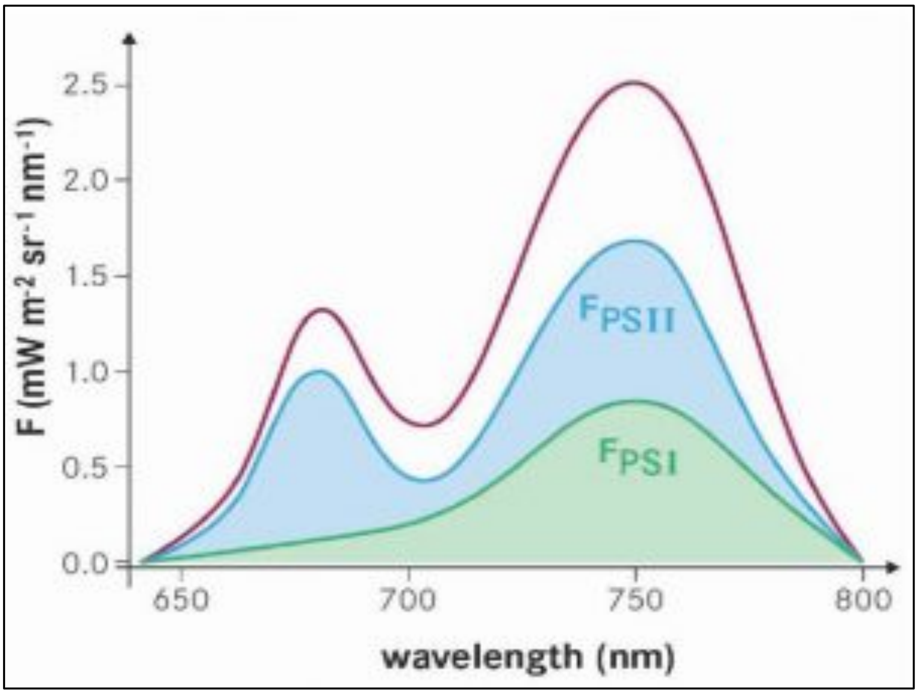


Figure 2 Fluorescence emission with typical contribution from both photosystem-1 (PS-1) and photosystem-2 (PS-2) (Report for Mission Selection, 2015)

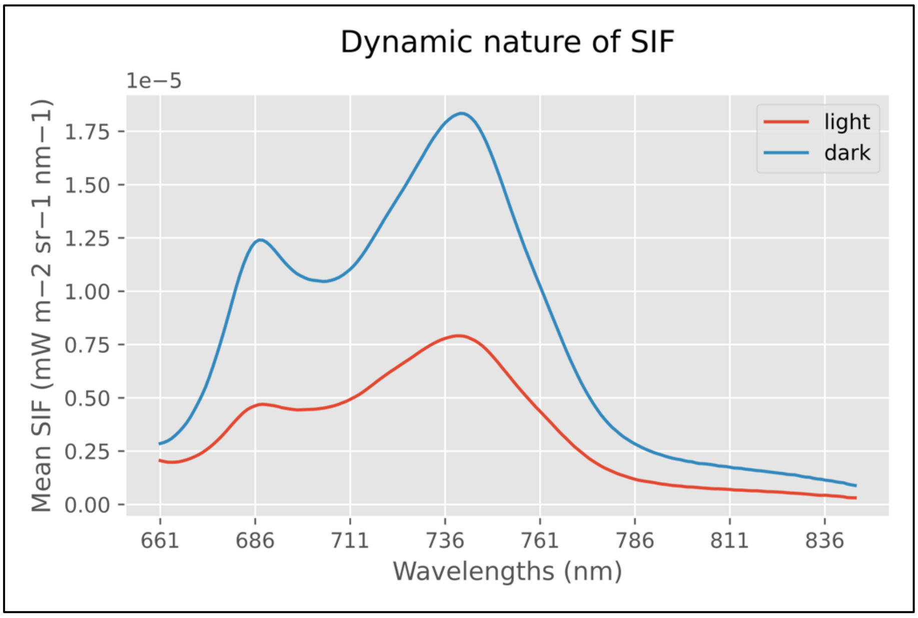


Figure 3 Dynamic nature of SIF in dark and light conditions

1.2 Literature Review

Plant biophysical variables such as pigment content, photosynthetic capacity, productivity, etc., can be quantified from spectral data using two approaches: 1) physically-based 2) statistical. Verrelst et al. (2019) mentions that physically-based approaches use physical laws that guide light interaction cause-effect relationships inside plants. These approaches use radiative transfer models (RTM) for inferring model variables. Such RTMs incorporate commonly accepted knowledge of physical principles. However, statistical approaches try to establish a direct relationship between target and input variables without relying on physical principles.

Several researchers have compared physical-based methods with statistical methods (Verrelst et al., 2019). They concluded that physical-based RTMs used with spectral data are commonly applicable and can inherently handle multicollinearity in input. However, they also added that this approach is computationally expensive. Additionally, RTM can map only its state variables as opposed to a wide range of biophysical variables. However, using statistical approaches, we can address the previous limitation.

The researchers mentioned that the statistical approaches are flexible in relating spectral data with quantifiable biophysical variables (Verrelst et al., 2019). Although these techniques suffer from multicollinearity, with subset selection and dimensionality reduction, multicollinearity can be alleviated. Statistical approaches are now widely used to infer plant properties from spectral data. Primarily, machine learning techniques have shown the potential to link plant-related biophysical variables and spectral data through adaptive and non-linear relationships.

Several researchers have attempted to predict plant properties from spectral data. Using spectral reflectance data, Shah et al. (2019) retrieved leaf chlorophyll content in wheat. They used a random forest machine learning model to link chlorophyll and reflectance. They obtained R^2 of 0.89 and root mean squared error (RMSE) of 5.49 $\mu g \cdot cm^{-2}$. Furthermore, An et al. (2020) used Gaussian process regression and random forest regression to predict rice chlorophyll content from hyperspectral reflectance data. They obtained R^2 of 0.80 and 0.76, respectively.

However, very few researchers have used spectral data to infer photosynthetic capacity using machine learning techniques. Fu et al. (2019) predicted photosynthetic capacity from hyperspectral leaf reflectance using stacked regression techniques. They obtained R^2 of 0.63 and RMSE of 36.4 µ mol m^{-2} s^{-1}. Zhou et al. (2021) used random forest model to estimate photosynthesis from hyperspectral reflectance data. The model obtained an R^2 of 0.92. These findings imply that machine learning techniques have the potential to link plant properties such as photosynthesis to hyperspectral data.

Nonetheless, there remains a research gap for predicting photosynthetic capacity from sun-induced fluorescence (SIF). Furthermore, the use of SIF to monitor photosynthesis is now widespread due to advancements in satellite measurements (Porcar-Castell et al., 2014). In this project, I decided to use less explored SIF data to predict photosynthesis using a popular machine learning technique.

1.3 Objectives and Scope

The project aims to develop a machine learning model to predict plant photosynthesis function using SIF data. Furthermore, I also hypothesize that a handful of SIF wavelengths can effectively predict photosynthesis function instead of using all possible SIF wavelengths. I divided the overall objectives into two parts:

1) *To use the random forest regressor model to estimate plant photosynthesis function from SIF data:*

   This objective helped determine whether SIF data can be used as a proxy of photosynthetic efficiency.

2) *To identify a small optimal number of bands from the full spectrum to effectively predict photosynthesis.*

The dataset contained a range of hyperspectral data, including leaf level, forest canopy level, and satellite level. However, only the leaf-level dataset was used in this project due to time constraints. Furthermore, the leaf level data contained several measurements of pigments (ancillary data), SIF for different wavelength bands, and reference photosynthetic efficiency measurements from an active fluorescence Pulse Amplitude Modulation (PAM) device. However, to limit the scope of this work, ancillary data was not considered. This is because the primary focus of the study was to relate SIF values for different bands to active fluorescence data from the PAM device.

2. Dataset

This chapter describes the dataset used in the study.

The leaf-level dataset was collected at CSIRO’s long-term forest research site called Tumbarumba, located in southern NSW (35° 39′ 20″ S, 148° 09′ 07″ E, 1200 m elevation). The data had 193 observations and 191 columns.

SIF measurements for band central wavelength (nm) (measured in mW m^{-2} sr^{-1} nm^{-1})							Photosynthetic function (unitless)
661	**662**	**663**	**…**	**848**	**849**	**850**	**YPSII**
4.44E-06	4.56E-06	4.73E-06	…	1.42E-06	1.34E-06	1.22E-06	0.699
3.48E-06	3.68E-06	3.90E-06	…	1.30E-06	1.22E-06	1.20E-06	0.737
2.96E-06	2.94E-06	2.94E-06	…	5.52E-07	5.65E-07	5.72E-07	0.629
2.46E-06	2.46E-06	2.45E-06	…	5.12E-07	5.01E-07	4.19E-07	0.656
2.62E-06	2.65E-06	2.71E-06	…	6.83E-07	7.09E-07	6.15E-07	0.726

Table 1 Preview of leaf level data

Table-1 shows the first five observations of the dataset. The data had SIF measurements for 190 wavelength bands. These wavelengths ranged from 661 nm to 850 nm with a spacing of 1nm. Unit of SIF is mW m^{-2} sr^{-1} nm^{-1}. These data were collected using an ASD Field Spectrometer (ASD Inc., Boulder, Colorado). The last column YPSII (Yield of photosystem-II), shows the photosynthetic efficiency of the plant. YPSII is unitless as it a fraction of absorbed energy available for photosynthesis for PSII. YPSII ranges between 0 and 1, where a healthy relaxed leaf has values around 0.8 and stressed leaves < 0.8. It was measured using a device called PAM (illustrated in figure-4).

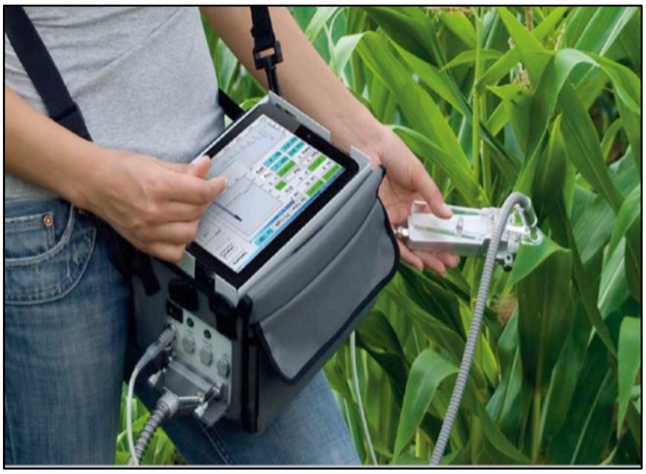


Figure 4 Pulse Amplitude Modulation (PAM) device to measure YPSII

3. Methodology

This chapter describes the adopted methodology to achieve the objectives. The methodology closely follows data-science approach. In this chapter section 3.1 describes the workflow for objective-1. Section 3.2 addresses the workflow of objective-2.

3.1 Objective-1

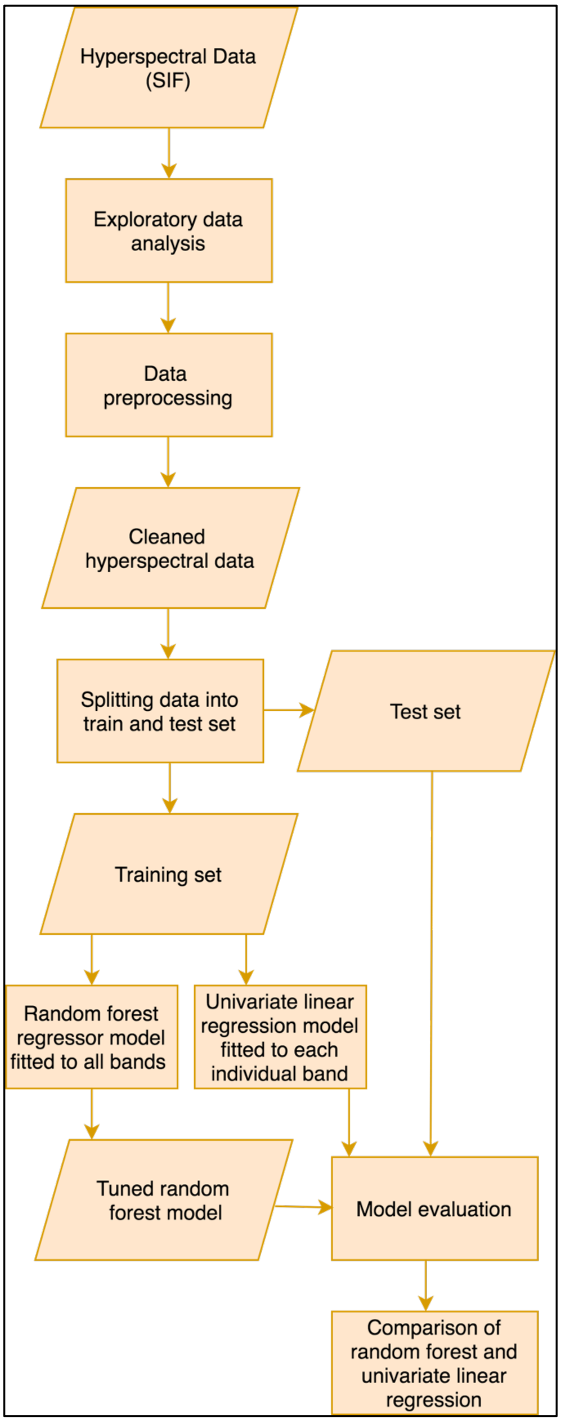

Hyperspectral Data (SIF)
Exploratory data analysis
Data preprocessing
Cleaned hyperspectral data
Splitting data into train and test set
Test set
Training set
Random forest regressor model fitted to all bands
Univariate linear regression model fitted to each individual band
Tuned random forest model
Model evaluation
Comparison of random forest and univariate linear regression

Figure 5 Workflow for objective-1

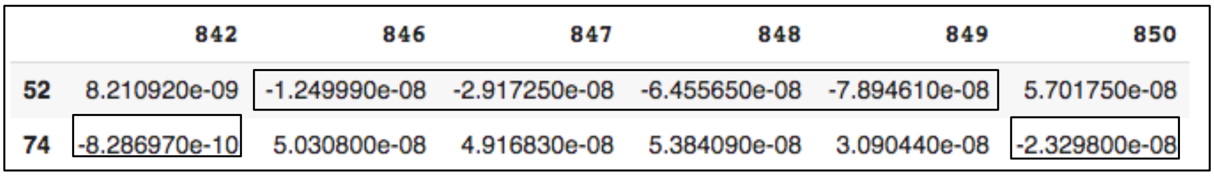
The objective-1 aimed at developing a random forest regressor model to accurately predict YPSII using SIF. The flowchart in figure-5 displays the workflow of objective-1. Briefly, I first performed Exploratory Data Analysis (EDA) to identify potential issues in the data. Then I preprocessed the data as guided by EDA findings. I split the ‘cleaned’ hyperspectral data into two sets – training and test set. Train set contained 80% of the data whereas test set used the rest 20% of the data. I attempted to maintain the distribution of YPSII in these subsets by taking uniform samples from each of the quartile regions. This was an essential step as the YPSII data was highly skewed (explained in section 3.1.1). The train data was then used to fit the random forest regressor model. The model was tuned using cross-validation over a parameter grid. I evaluated the model performance on the held-out test on various evaluation metrics. As a benchmark model, I also trained univariate linear regression on each band to compare its performance with the random forest model.

In this section I will explain the steps involved in objective-1 workflow.

3.1.1 Exploratory Data Analysis (EDA)

EDA is performed to understand the data's characteristics and to find out issues in the data, often employing summary statistics and data visualization methods.

1) Bands with negative values



	842	846	847	848	849	850
52	8.210920e-09	-1.249990e-08	-2.917250e-08	-6.455650e-08	-7.894610e-08	5.701750e-08
74	-8.286970e-10	5.030800e-08	4.916830e-08	5.384090e-08	3.090440e-08	-2.329800e-08

Figure 6 Negative SIF values in the data

Figure-6 highlights negative SIF values. It can be observed that bands 842 and 850 have negative values in row 74, whereas bands 846, 847, 848, and 849 have negative values in row 52. These bands were not considered for the modeling because negative values are not valid and are caused by low signal and instrument noise.

2) Multicollinearity

Multicollinearity occurs when the columns are highly correlated with each other (Frost, 2017). The correlation heatmap shown in figure-7 indicates the wavelengths are highly correlated with each other. Furthermore, the minimum correlation among any two columns was 0.81.

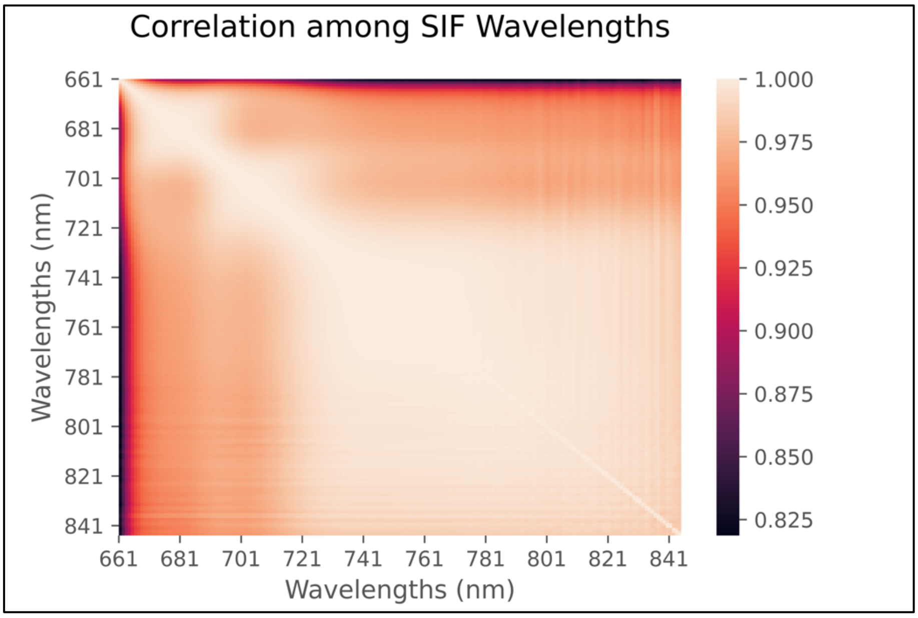


Figure 7 Heatmap illustrating multicollineaity in SIF wavelengths

Linear models such as linear regression, support vector machines are affected by the presence of multicollinearity. However, ensemble methods such as random forest are robust to multicollinearity because of their bootstrap feature sampling (Strobl et al., 2008).

3) Distribution of target variable(YPSII)

Figure-8 shows the distribution of YPSII using a density plot. YPSII is negatively skewed and the density of YPSII is higher near 0.7. Since many observations have larger YPSII, it can introduce a bias towards higher YPSII in machine learning model predictions. However, this can be addressed by discretizing the YPSII column.

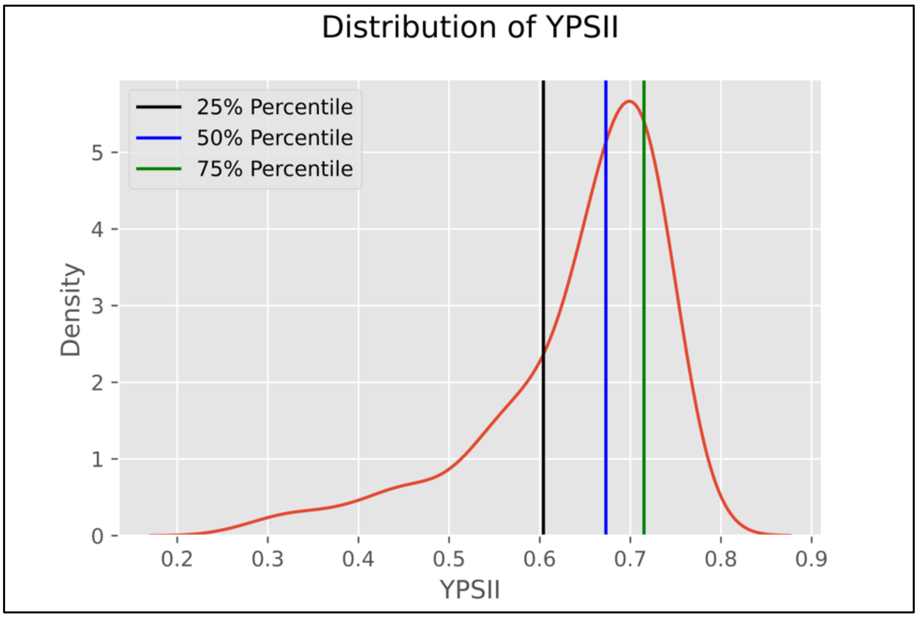


Figure 8 Density plot of target variable (YPSII) with its quartiles

I discretized the column YPSII based on the quartiles, i.e., 25th, 50th, and 75th percentiles, to address the skewness problem. Mainly, I assigned classes to the observations based on the quartile region they fell into. The following table-2 illustrates the discretization process.

Region	Class	Number of observations in the class
0 < YPSII < 25th percentile	1	48
25th percentile < YPSII < 50th percentile	2	48
50th percentile < YPSII < 75th percentile	3	48
75th < YPSII < 1	4	49

Table 2 YPSII discretization

After discretization, each class contained a similar number of observations. The discretization facilitated stratified sampling for machine learning model training and testing dataset. Furthermore, discretization also maintained YPSII distribution in the training and testing datasets.

3.1.2 Random Forest Regressor

Random Forest (RF) is an ensemble machine learning model which uses many decision trees as a base estimator, and the model outcome is obtained by aggregating the output of individual decision trees (Breiman, 2001). Though many different machine learning models can be fitted to spectral data, RF has shown robustness to overfitting (An et al., 2020; Shah et al., 2019). RF randomly selects a subset of predictors to train the individual decision trees, makes it an accurate model (Breiman, 2001).

Furthermore, RF also provides insights into each predictor variable's importance, making it useful for inference (Breiman, 2001). The regression version of RF is a random forest regressor.

It is crucial to understand how decision trees are fitted to understand how RF works. Decision trees are non-linear machine learning models that use binary decision rules to determine the final output. An illustration of a simple decision tree is displayed in figure-9. Each tree starts with a root node, extends further through internal nodes, and terminates at the leaf node (Breiman, 2001). The splitting of nodes are usually decided using Mean Squared Error (MSE) criterion (Georgios Drakos, 2019) for regression problems.

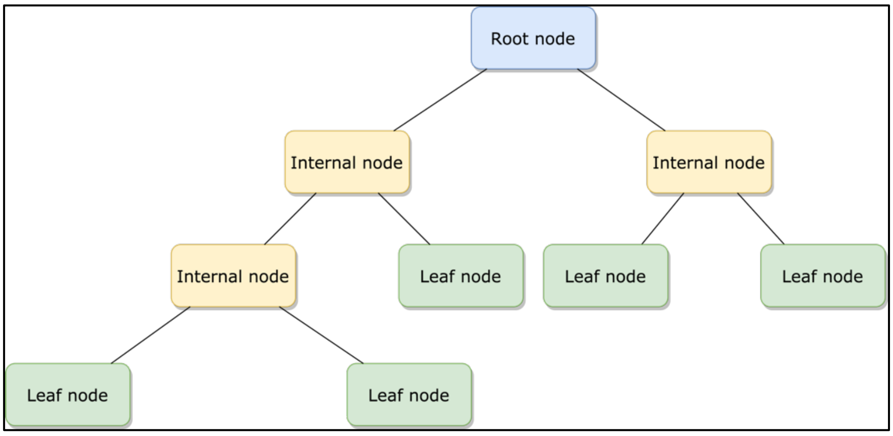


Figure 9 Illustration of decision tree

Mathematical formulation of decision trees for regression

The following mathematical formulation is adapted from scikit learn decision tree user guide. (*1.10. Decision Trees — Scikit-Learn 0.24.2 Documentation*, n.d.)

Consider training set vectors $x_i \epsilon \mathrm{R}^{\mathrm{n}}$ where $i = 1, \ldots, N$ and target (label) values $y_i \epsilon \mathrm{R}$, decision tree regressor splits the feature (predictor) space in a way that similar labels belong to the same split.

Consider for any node p (illustrated in figure-9) , the data at the node is S_p with N_p data points. The decision tree splits the feature space and choose the best split by evaluating its quality on loss function. Each candidate split $\theta = (\mathrm{j}, \mathrm{t_p})$ where j is the feature to split given threshold t_p, divides data into two partitions $S_p{}^{left}(\theta)$ and $S_p{}^{right}(\theta)$.

$$S_p{}^{left}(\theta) = \{(x, y) | x_j \leq t_p\}$$
$$\mathrm{S_p}{}^{\mathrm{right}}\,(\theta) = \mathrm{S_p}/\mathrm{S_p}{}^{\mathrm{right}}(\theta)$$

Regression decision tree uses MSE as a loss function $L()$ to evaluate the quality of the split. The loss calculated at node p with split θ is given by $G\left(S_p, \theta\right)$.

$$\mathrm{G}\left(S_p, \theta\right) = \frac{N_p{}^{left}}{N_p} \,\mathrm{L}\left(S_p{}^{left}(\theta)\right) + \frac{N_p{}^{right}}{N_p} \,\mathrm{L}\left(S_p{}^{right}(\theta)\right)$$

Where $\mathrm{L}\left(\mathrm{S_p}(\theta)\right) = \frac{1}{\mathrm{N_p}} \sum_{\mathrm{y} \epsilon \mathrm{S_p}} \left(\mathrm{y} - \bar{y}_p\right)^2$ and $\bar{y}_p = \frac{1}{N_p} \sum_{y \epsilon S_p} y$

The selected split is $\theta^* = \mathrm{argmin}_\theta\ \mathrm{G}\left(\mathrm{S_p}, \theta\right)$. The recursive process of selecting subsets $S_p{}^{left}(\theta)$ and $S_p{}^{right}(\theta)$ for different nodes is performed until $N_p = 1$ or N_p reaches desired data points.

Random Forest Regressor Explanation

RF uses many such decision trees and takes the average of their output to determine the outcome, as illustrated in figure-10. The study supplied three parameters to the random forest regressor – i) number of trees (n_estimators), ii) maximum depth of individual trees (max_depth), and iii) minimum sample for a leaf node (min_sample_leaf). The number of trees specifies the number of individual decision trees used to make a forest. The maximum depth of a tree controls how much a tree should be allowed to expand. The final parameter specifies a minimum number of observations required to form a leaf node. These parameters were optimized using cross-validation over a parameter grid (explained in sections 3.1.3 and 3.1.4).

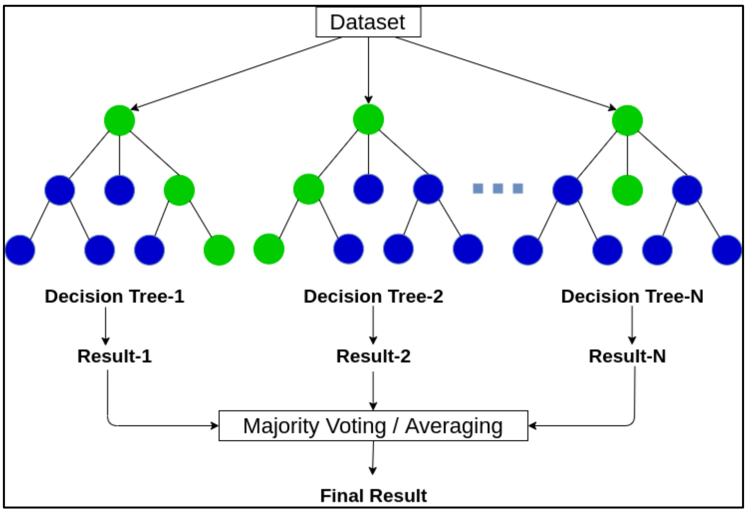


Figure 10 Random forest as an ensemble of many decision trees (Abhishek Sharma, 2020)

3.1.3 Grid-Search

Grid-search is a parameter tuning algorithm, used to select the optimal parameter-set from the given grid of parameters (Lutins, 2017). I used grid search with cross-validation (explained in section 3.1.4) to choose optimal parameters for the model. Table-3 displays the parameter grid for the model. Particularly, the algorithm exhaustively searches the best parameter combination from possible combinations of the parameters. It uses 5-fold cross-validation to evaluate model performance trained using a combination of parameters. The selected parameter-set has the best model performance in a 5-fold cross-validation.

Model parameter	Set of parameter values
Number of estimators	{100, 200, 300, 500, 1000}
Maximum depth	{10, 20, 50, 100, 200}
Minimum samples for leaf node	{1, 5, 10, 20}

Table 3 Parameter grid of the random forest model

3.1.4 Cross Validation

In machine learning, the dataset is usually divided into three parts – train, validation, and test. However, in the case of a smaller dataset, these partitions would reduce the data available for the model to train. Hence, a typical practice is to use cross-validation instead separate validation set when the dataset contains fewer samples.

Cross-validation is a statistical method for evaluating model performance during the training stage (Leite, 2020). It can also be used to select optimal parameters for a model. I used 5-fold cross-validation to identify the best parameters. Particularly, I divided the training dataset into 5-folds as shown in the figure-11. Each fold contained 30 or 31 data points. I trained the model on four folds for a particular combination of parameters from the table-3 while leaving out one fold. The trained model was evaluated on the held-out fold. This process was repeated until each fold is once used as a held-out fold. The final model performance is a mean performance of the model on each held-out fold for a given combination of parameter-set. I selected the parameter combination which has the best performance in 5-fold cross-validation.

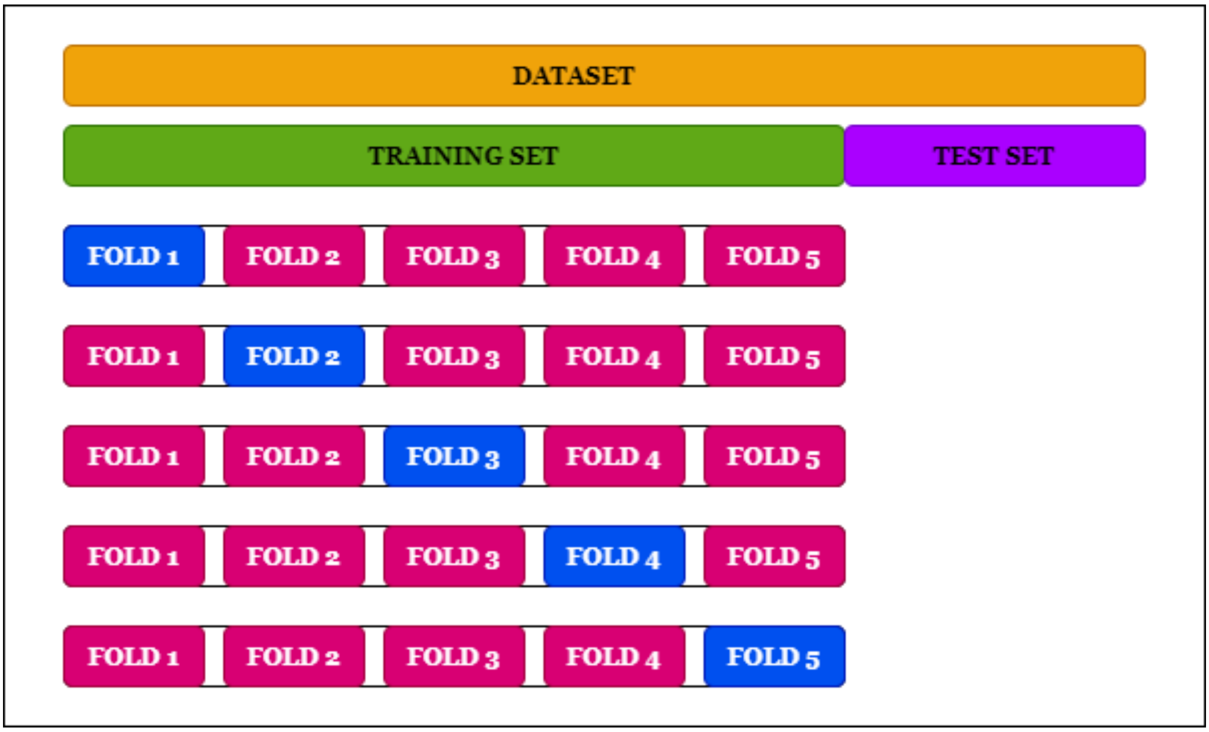


Figure 11 Illustration of 5-fold cross-validation (Leite, 2020)

3.1.5 Model Evaluation

The model selected from grid-search cross validation was evaluated on a separate test-set. I used Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-Squared (R^2) as evaluation metrics. The relevant mathematical basis is explained below (Wu, 2020).

1) Mean Squared Error (MSE)

MSE is calculated as mean of the squared differences between ground truth of the target variable and predicted values of the target variable using the model. For a dataset with N samples with target variable y and model predictions $\hat{y}$, then MSE can be calculated using following formula.

$$\mathrm{MSE} = \frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2$$

2) Mean Absolute Error (MAE)

MAE is calculated as a mean of the absolute differences between ground truth of the target variable and model predicted values of the target variable. For a dataset with N samples with target variable y and model predictions $\hat{y}$, then MAE can be calculated using following formula.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_\mathrm{i}|$$

3) R-Squared (R^2)

R^2 value indicates how much variability in the target variable is explained by a given set of predictors using the model. It can be calculated by subtracting ratio of sum of squares of residuals to total sum of squares from unity. For a dataset with N samples with target variable y, mean target value $\bar{y}$ and model predictions $\hat{y}$, then R^2 can be calculated using the following formula.

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} i$$

$$\mathrm{S}S_{residuals} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\mathrm{S}S_{total} = \sum_{i=1}^{N}(y_i - \bar{y}_i)^2$$

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

3.2 Objective-2

The second objective aimed at selecting a small number of useful SIF wavelength bands that can effectively predict YPSII.

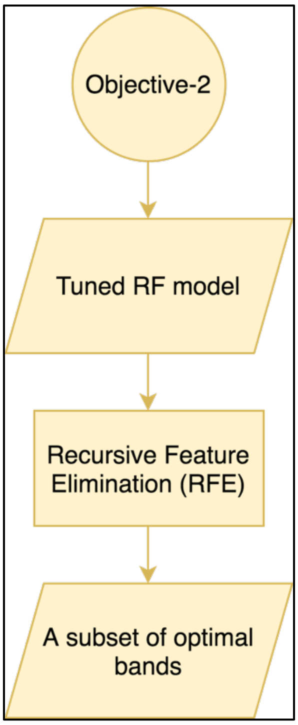

Objective-2
Tuned RF model
Recursive Feature Elimination (RFE)
A subset of optimal bands

Figure 12 Proposed workflow for objective-2

Figure-12 illustrates workflow for the second objective. I used the previously tuned model to identify optimal number of bands using Recursive Feature Elimination (RFE). The following section explains the mathematical basis of RFE.

3.2.1 Recursive Feature Elimination (RFE)

RFE is a commonly used feature (predictor) selection technique in machine learning. It selects important features by recursively reducing the size of the input feature set (Brownlee, 2020). Figure-13 displays a flowchart of the RFE method adapted to the project. The method required four inputs which include the tuned model, initial number of features, number of features to select, and number of features to remove at each step. The model was initially trained with the initial number of features using a cross-validation approach. At each step, based on feature importance score from a model, a fixed proportion of features were removed. This step led to a reduced subset of features. The model was again trained with this subset using cross-validation. This process was repeated until the subset size reached a pre-defined subset size.

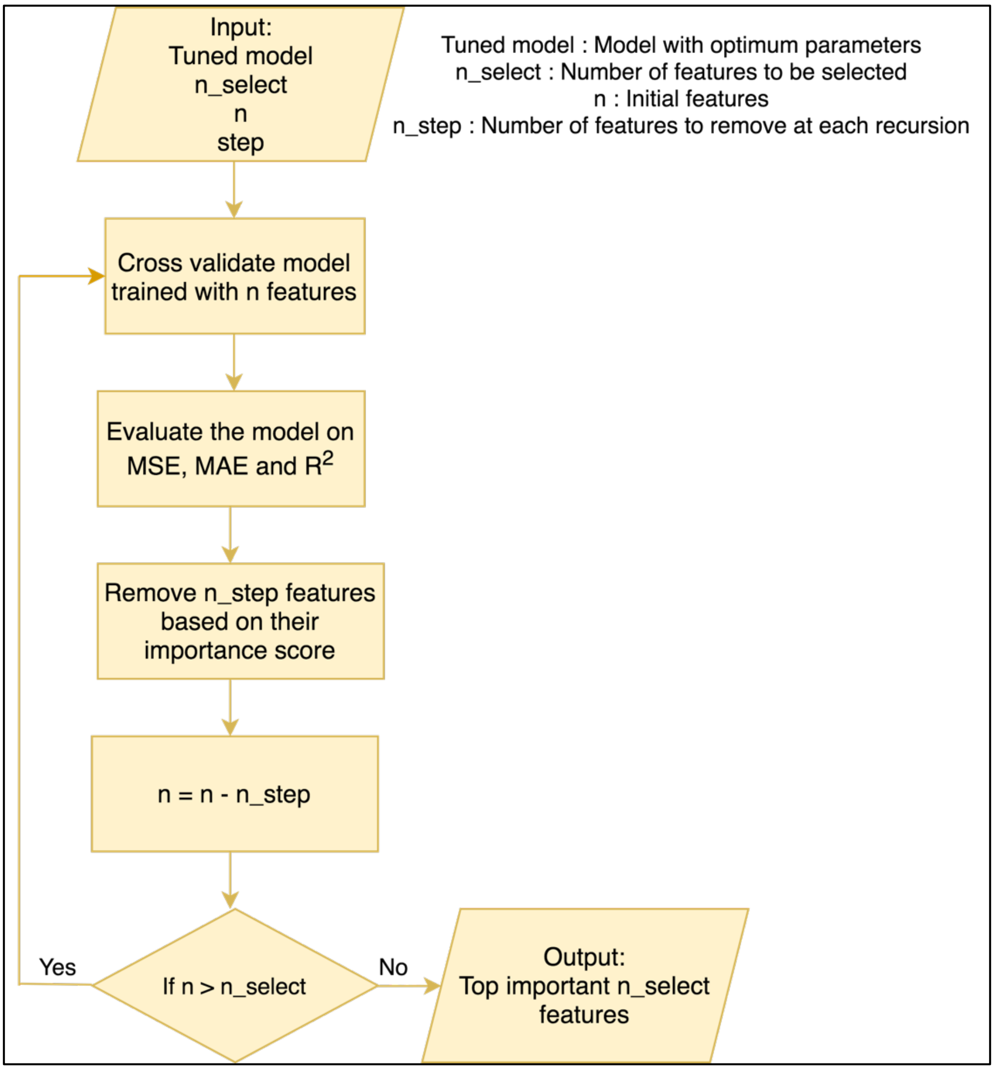

Input:
Tuned model
n_select
n
step
Tuned model : Model with optimum parameters
n_select : Number of features to be selected
n : Initial features
n_step : Number of features to remove at each recursion
Cross validate model trained with n features
Evaluate the model on MSE, MAE and R2
Remove n_step features based on their importance score
n = n - n_step
If n > n_select
Yes
No
Output:
Top important n_select features

Figure 13 Illustrating Recursive Feature Elimination (RFE)

4. Results

This section describes important results from each objective.

4.1 Objective-1

The purpose of the objective was to examine whether a machine learning approach such as Random Forest Regressor can accurately predict YPSII values using all SIF bands. This section presents the result of the objective.

4.1.1 Univariate Linear Regression

To understand whether each SIF band can be used to predict YPSII, I performed regression between individual SIF bands and YPSII. The model was evaluated on the test set using MSE, MAE, and R^2.

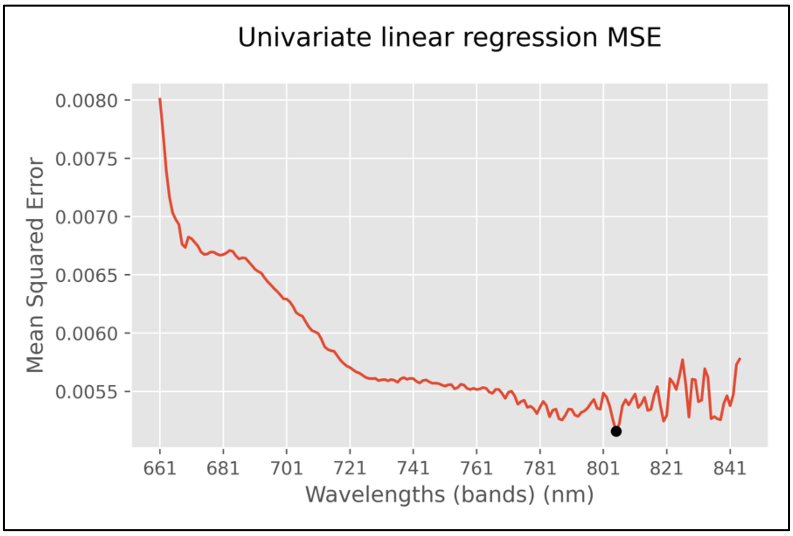


Figure 14 MSE calculated for univariate linear regression on test set. The lowest MSE was 0.0051 for wavelength 805 nm.

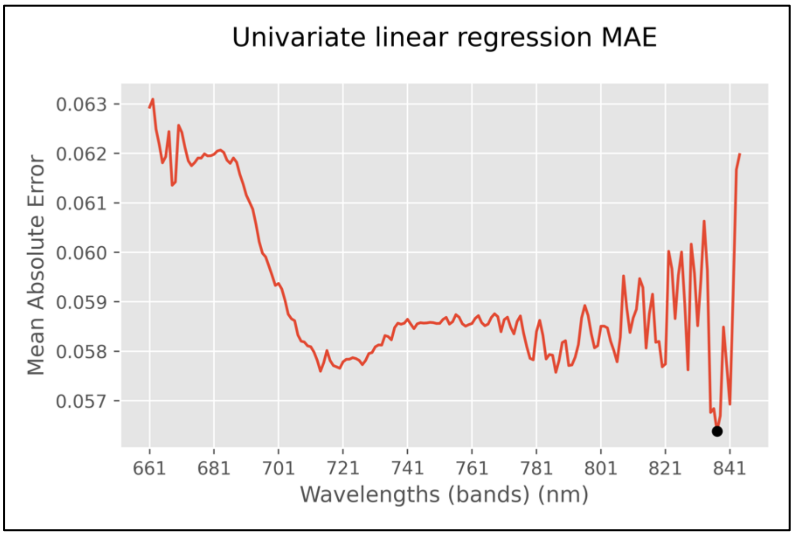


Figure 15 MAE calculated for univariate linear regression on test set. The lowest MAE was 0.0563 for wavelength 837 nm.

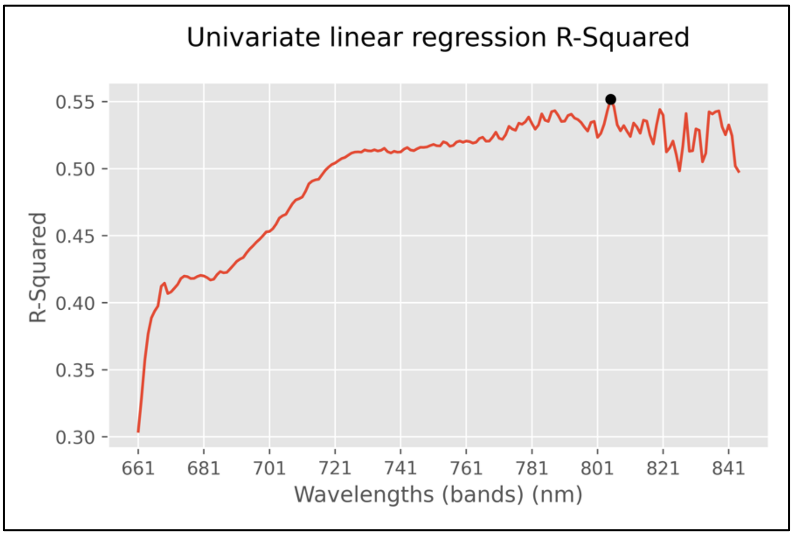


Figure 16 R-Squared calculated for univariate linear regression on test set. The highest R^2 was 0.55 for wavelength 805 nm.

Figure-14 to 16 display MSE, MAE and R^2 plotted against individual wavelengths. It can be observed that the highest R^2 obtained from univariate linear regression was 0.55, indicating 55% of the variance in YPSII can be explained by a single wavelength 805 nm. Furthermore, I also obtained the minimum MSE for the same wavelength, 805. Though, the minimum MAE 0.0563 was obtained for wavelength 837. However, investigation of full-spectrum range instead of individual wavelengths using the RF model will improve the predictive accuracy of YPSII.

4.1.2 RF using all spectral bands

I trained the RF model with all available bands to see the synergistic contribution of the full spectrum of bands in predicting YPSII. The model was evaluated on the test set with MSE, MAE, and R^2.

Evaluation metric	Univariate linear regression	Random forest model
MSE	0.0051	0.0016
MAE	0.0563	0.0292
R^2	0.5500	0.8604

Table 4 Comparing univariate linear regression and random forest model performance on test set

Figure-17 shows a 95% prediction interval of model predictions on the test set. Almost all ground truth values of YPSII fell in the 95% prediction interval bounds, confirming RF as an accurate estimator of YPSII. Table-4 compares the random forest model with univariate linear regression. The random forest model trained with all bands performed much better than the univariate linear regression model trained with individual bands.

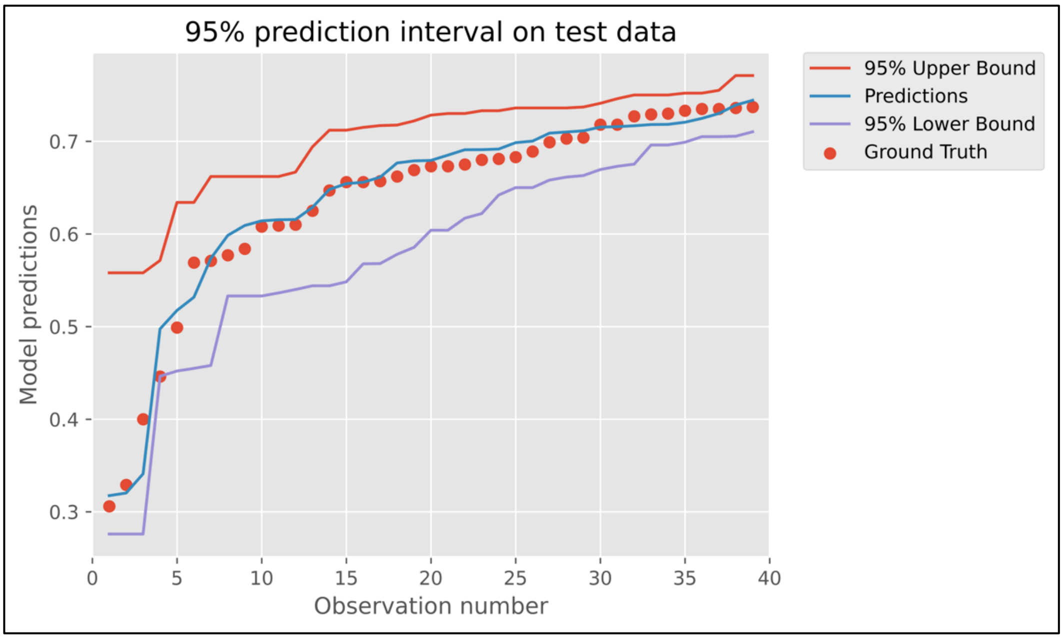


Figure 17 Plot of 95% prediction interval for the random forest model on test set predictions

4.1.3 Optimization of RF

RF model optimization was an essential part of the model training due to many tunable model parameters. As described earlier, the study used cross-validation over a grid of parameters to choose optimal parameter values. The two most important parameters were min_sample_leaf (minimum samples required for a leaf node) and n_estimators (number of trees). Min_sample_leaf controls the depth of a tree, and n_estimators controls how dense a forest is. The balance between them is essential for accurate predictions. Figure-18 shows a relation between the two parameters. It can be observed that for a given number of trees, as min_sample_leaf increased, MSE also increased. This implies that lower min_sample_leaf was needed for more accurate predictions. Furthermore, figure-19 shows that minimum MSE was obtained for 100 trees. So the chosen parameters were 100 trees and one min_sample_leaf.

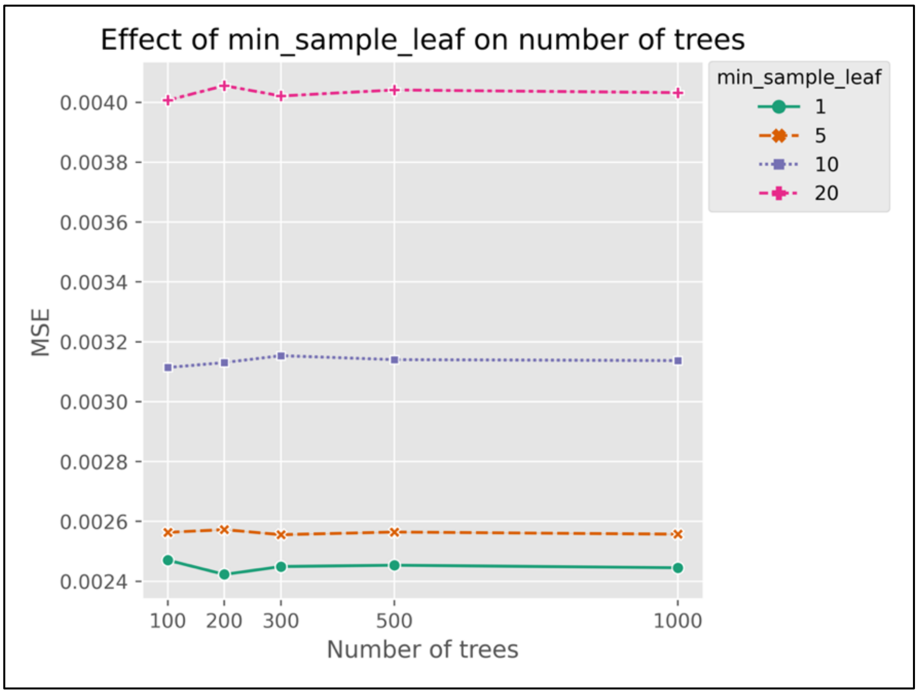


Figure 18 MSE values for various min_sample_leaf and number of trees.

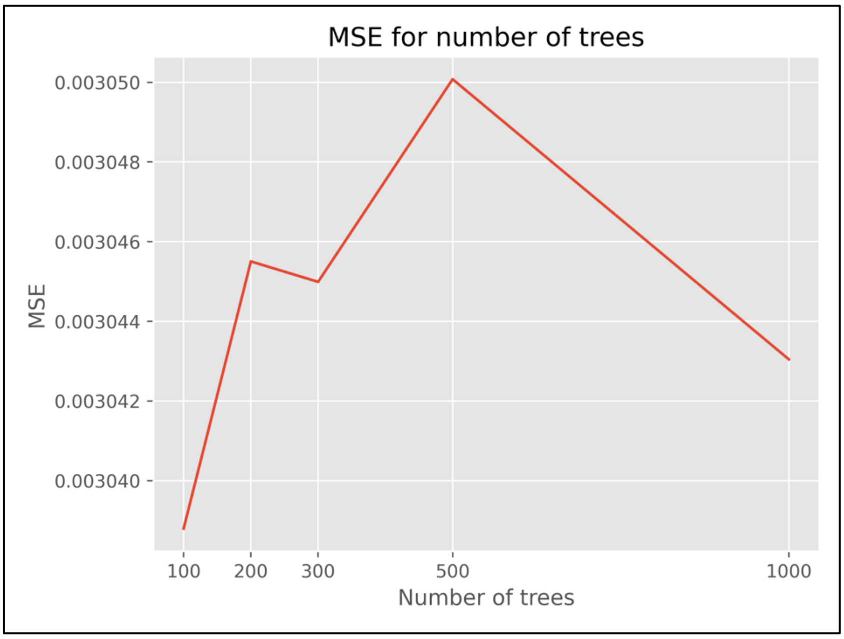


Figure 19 Effect of number of trees on MSE

4.1.4 Random forest model importance

The RF model can also be used to identify essential features (predictors). I identified essential bands from the tuned random forest model. I considered both cases where I trained the tuned model on 500 bootstrapped samples and one without bootstrapping. Figure-20 shows the importance of each wavelength as identified by RF model without using bootstrapped samples. Figure-21 shows wavelength importance from the model trained with bootstrapped samples. SIF spectrum is also included in the figures to compare the importance of each wavelength with its SIF signal.

Two key observations can be made from these plots.

1) The wavelengths corresponding to two peaks in the SIF spectrum has relatively higher importance scores.

2) Not all wavelengths are equally contributing to predict YPSII. It can be observed that several wavelengths' importance score is close to zero. This implied YPSII can be predicted using a smaller set of wavelengths instead of a full spectrum.

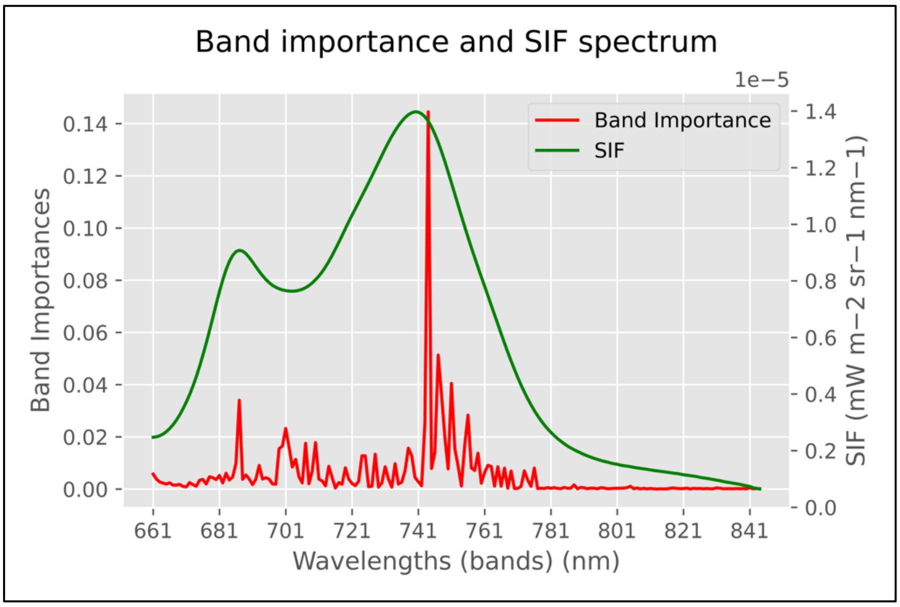


Figure 20 Importance of each band plotted along with SIF spectrum (without bootstrapping)

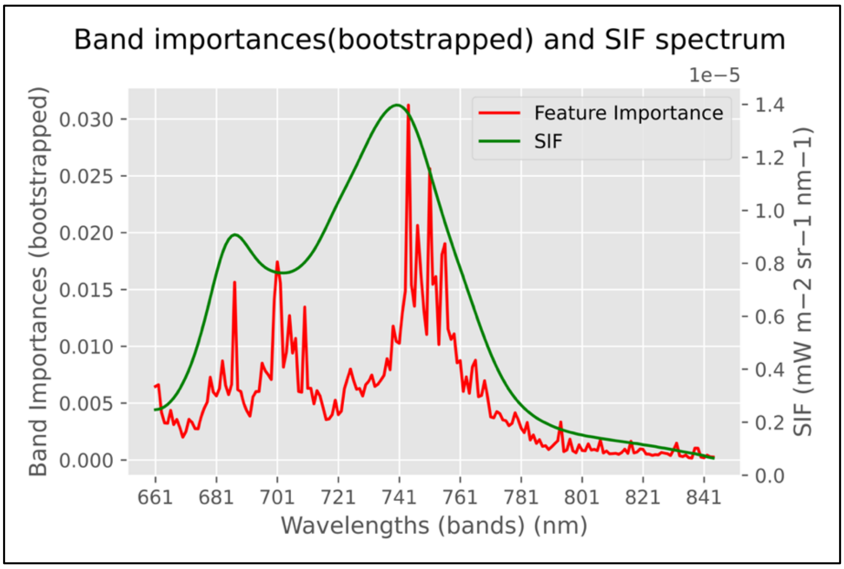


Figure 21 Importance of each band plotted along with SIF spectrum (with bootstrapping)

4.2 Objective-2

The purpose of this objective was to find optimal bands from the entire SIF spectrum, which can accurately predict YPSII. The following section describes results of each step in objective-2.

4.2.1 Selection of optimal features

Figures 22-24 illustrate MSE, R^2 and MAE values plotted against various bands used in model training. It can be observed that a sharp decline in MSE/MAE until approximately 40 wavelengths; after that, the reduction became slower. This implies 40 was an optimal number of bands as opposed to using the full 184 bands.

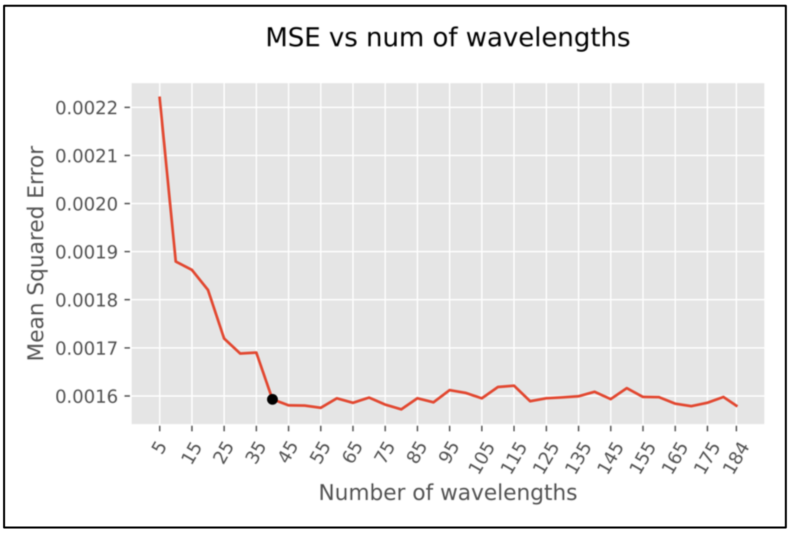


Figure 22 MSE plotted against different number of wavelengths used in model training. The optimal number of wavelengths was 40 as shown by elbow point.

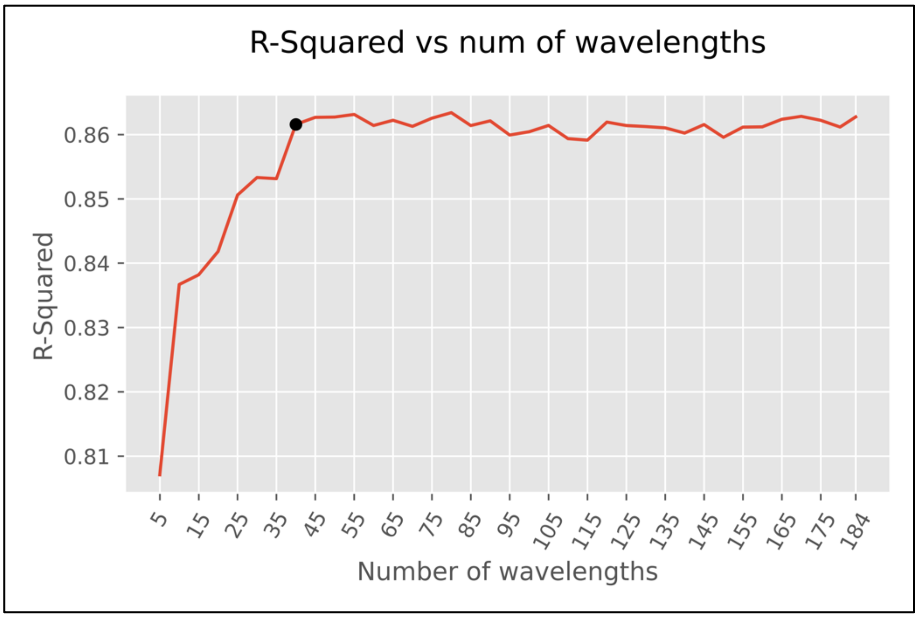


Figure 23 R-squared values plotted against different number of wavelengths used in model training. The optimal number of wavelengths was 40 as shown by the knee point.

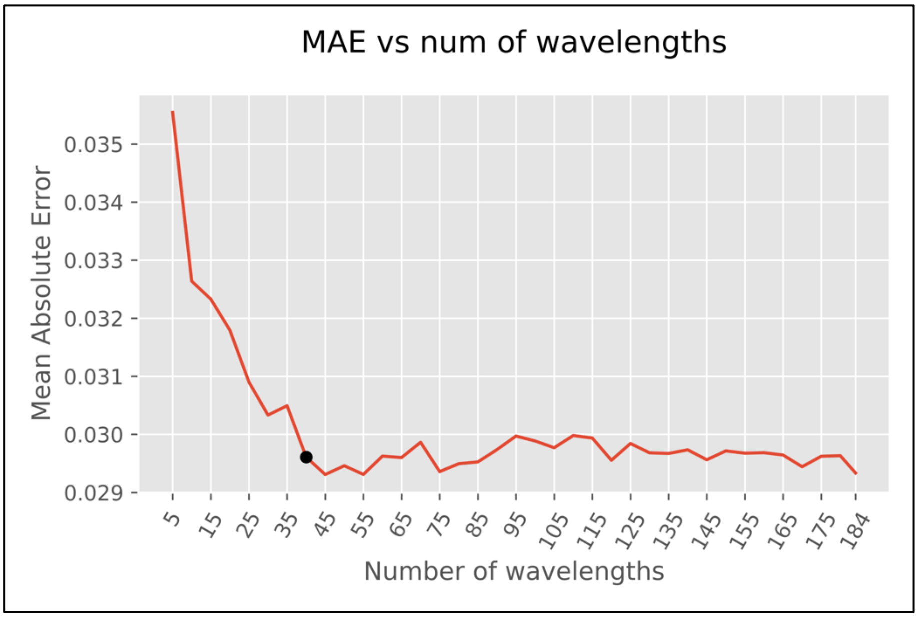


Figure 24 MAE plotted against different number of wavelengths used in model training. The optimal number of wavelengths was 40 as shown by elbow point.

The optimal 40 spectral bands are highlighted in figure-25 along with SIF spectrum. Table-5 contains these 40 bands. Again, they coincide with the two peaks of the SIF spectrum as observed in figure 20-21.

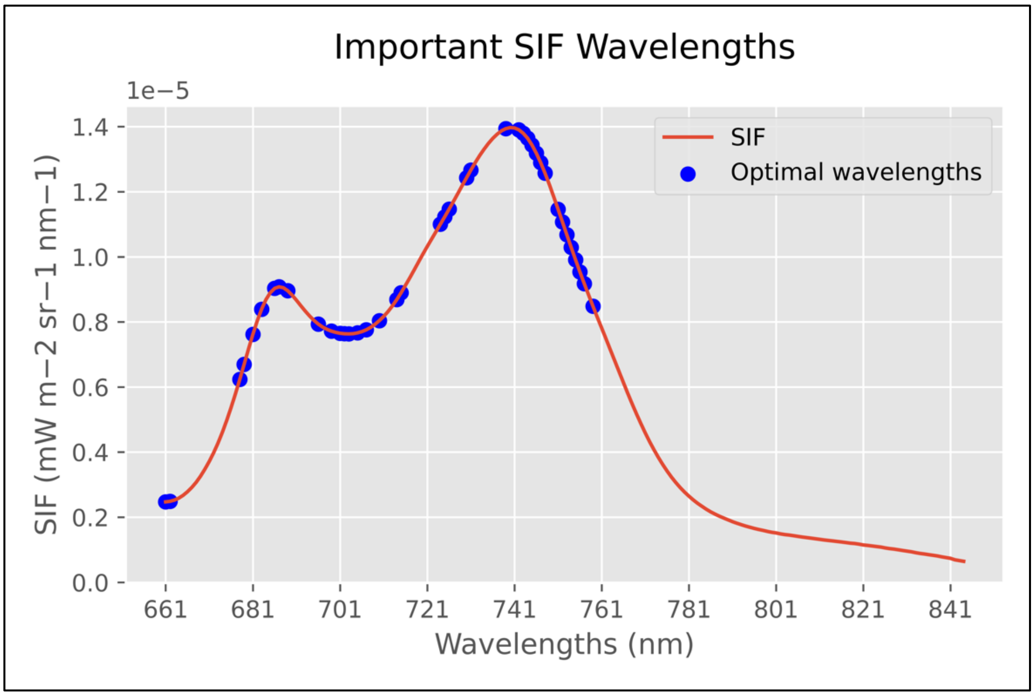


Figure 25 - Top 40 optimal SIF wavelengths

40 Most Important SIF Bands (nm)									
661	662	678	679	681	683	686	687	689	696
699	701	702	703	705	707	710	714	715	724
725	726	730	731	739	742	743	744	745	746
747	748	751	752	753	754	755	756	757	759

Table 5 Top 40 important SIF bands

4.2.2 Comparison of optimal wavelengths with full spectrum of wavelengths

Table-6 shows performance of the RF model trained with the extracted 40 bands against full bands It is evident that model with 40 bands has a similar performance to the model with full bands.

Evaluation metric	Random forest regressor with 40 bands	Random forest regressor with full bands
MSE	0.0296	0.0292
MAE	0.0015	0.0016
R^2	0.8615	0.8604

Table 6 Model performance comparison for random forest regressor model trained with 40 bands and with full bands

5. Discussion

In this section I will compare the alternative machine learning approaches and justify my decision of choosing random forest regressor. Furthermore, I will also discuss the results of the study.

5.1 Alternative Machine learning approaches and its evaluation

In the study, I used random forest regressor model among other models including Decision tree, artificial neural network and Support Vector Machines (SVM),. The strengths and weaknesses of each model are explained below:

1) **Decision Tree (Gupta, 2020):**

   Decision tree uses simple binary rules to make decision.

   Strengths:

   - It does not require data to be normalised or standardized. Hence, scale of a data usually does not affect the model performance.
   - Missing values in data do not have significant impact on model predictions.
   - It is interpretative and hence decision making process can be explained to other stakeholders.
   - Since it chooses features that significantly impact the target variable, irrelevant feature will not have considerable effect on model performance.

   Weaknesses:

   - It is vulnerable to overfitting as decision tree depth increases.
   - It is highly sensitive to data and hence outcome can significantly change as data changes.

2) **Random Forests (Gupta, 2020):**

   Random forests are ensemble of many decision trees.

   Strengths:

   - It is robust to overfitting due to ensemble nature of the model.
   - It provides higher accuracy as the outcome is decided by aggregating outcomes of the many decision trees.
   - It can handle imbalance dataset.
   - It is not highly susceptible to multicollinearity as each decision tree in the forest uses a different subset of features for model training.
   - It is useful to extract important features for inference.

Weaknesses:

- Less interpretative as it is difficult to know how each decision tree is making decisions.

3) **Artificial Neural Network (ANN)**

ANN is a powerful machine learning model that uses collection of connected nodes to make Decision.

Strengths: (Mahanta, 2017)

- It is capable of learning non-linear and complex relationship between predictors and target.
- It does not make assumptions on input data and hence very flexible.

Weaknesses: (Tu, 1996)

- It requires large amount of data for model training.
- Computationally expensive due to long training time.
- It is black-box in nature as it is difficult to derive how it made a decision.
- It is prone to overfitting as it uses many parameters to make decision.

4) **Support Vector Machine (SVM):**

SVM learns nonlinear functions by transforming features into higher-dimensional space using kernel functions (Verrelst et al., 2019).

Strengths: (Gupta, 2020)

- It performs well with high-dimensional data
- It is less susceptible to outliers

Weaknesses: (Gupta, 2020)

- It is important to choose the optimum parameter for the model
- Normalization or standardization is required to scale the data
- It is not interpretative and difficult to extract feature information

The dataset used for the study was not rich and hence models like SVM, ANN, decision tree could easily overfit. Furthermore, using SVM and ANN it is difficult to extract feature importance which was essential for my objective-2. Although the above limitations could be addressed with appropriate techniques but due to time-constraints I decided to use random forest regressor.

5.2 Discussion of Results

In this study, I investigated hyperspectral SIF data for the purpose of predicting photosynthetic capacity. Random forest regressor model was trained to predict YPSII from SIF data - 1) using all bands 2) using fewer optimal bands. For both the cases, model obtained similar results i.e. R^2 of 0.86. Since SIF data has not previously been much explored by researchers, it is difficult to make one to one comparison. However, several researchers have explored hyperspectral reflectance data to predict photosynthesis. The results of the study can be compared with previous results with reflectance data. Previously using stacked regression methods with hyperspectral reflectance data, the R^2 of 0.63 was obtained in predicting photosynthetic capacity (Fu et al., 2019). However, the random forest model used in my study greatly outperforms the stacked regression model on the reflectance data.

Furthermore, figure-21 illustrates that the importance of each band closely follows SIF spectrum i.e. the SIF signal is higher near two peaks and the importance of bands close to these peaks are also higher. Additionally these peaks are corresponding to two photosystems (figure-2), implies that SIF bands played an important role in predicting photosynthesis capacity of the photosystems.

6. Reflections

In this chapter, I will reflect on the deviation of the project from the project proposal. I will also mention the tools used in the project.

As per the proposal, the project aimed to develop predictive machine learning models at three different levels of data – leaf-level, forest-canopy level, and satellite level. However, I could only work on leaf-level data. This is mainly attributed to significant time spent in model tuning and training. I used the grid-search cross-validation approach to tune our model parameters. Since the grid search looks for the best parameter-set over the parameter space, it was computationally costly. Furthermore, I also used recursive feature elimination to select the optimal number of features, which was also computationally expensive. Due to these limitations, I decided to reduce the scope of our project to leaf-level data.

In the proposal, I proposed to implement various machine learning techniques. However, I focused on the random forest approach due to time constraints. This is mainly because other methods, such as support vector machines, neural networks, etc., cannot give insights into important features. The random forest can specify the importance of each feature used in modeling. This was particularly important to determine important SIF wavelengths for predicting YPSII.

I mainly used four project management tools during the project. For workplace communication, I used Slack (Slack, n.d.). We had a range of datasets, including SIF, reflectance, and transmittance. The data were shared through the University of Queensland's Research Data Manager(UQRDM) platform. Programming was performed in Python (*Welcome to Python.Org*, n.d.). I used python packages such as Sci-kit learn, Pandas, NumPy, MatPlotLib, and Seaborn for data preprocessing, machine learning, and data visualization. The programming and model building were mainly performed in Google's Colab (*Google Colaboratory*, n.d.).

7. Conclusions

Plants play an essential role in alleviating two major global issues, including food demand and global warming. Hence, it is imperative to maintain plant health. One way to measure plant health is to track the photosynthesis process. In this study, I prepared a machine learning model to predict plant photosynthetic capacity using SIF data. I also extracted 40 important SIF wavelengths, which acted as critical predictors of YPSII.

Particularly, with a random forest regressor, I obtained 86% of R^2 in predicting YPSII from SIF. The R^2 was much higher than the benchmark R^2 of 55% from univariate linear regression. Furthermore, I also extracted the importance of each band in predicting YPSII. Additionally, I also found that YPSII can be predicted from fewer bands instead of using the entire spectrum. The optimal bands were mainly corresponding to the two peaks in the SIF spectrum. The study can act as a base for further research on linking SIF and photosynthesis using machine learning techniques.

8. Recommendations

This section presents recommendations for future research activity based on the project findings and limitations. These mainly include using different machine learning models, enriching dataset, and potential industrial product development.

Although the study could obtain a highly accurate model for predicting YPSII using SIF wavelengths, there is still room for model improvement. First, the study was limited by the sample size of the input dataset. Therefore, I recommend using a large dataset for robust model building and evaluation, such as more plant species. Furthermore, the study only used random forest regressors. I, therefore, recommend using other popular machine learning models such as neural networks, support vector machines, Gaussian process regression, xgboost, etc. Additionally, stacked regression of these models can also be performed to overcome the limitation of each model (Fu et al., 2019).

Furthermore, the study mainly focused on leaf-level SIF data. However, the areal coverage of leaf-level data is much less as compared to tower-level and satellite-level data. Hence, using new technologies such as proximal sensing systems, SIF data can be collected at a bigger scale (William Woodgate et al., 2020). I therefore recommend extending the scope of machine learning modeling at a larger scale with the data collected using these technologies

Finally, as photosynthesis can be helpful in determining plant health (W. Woodgate et al., 2019), industrial products for real-time plant health monitoring can be developed.

9. Bibliography

1.10. Decision Trees—Scikit-learn 0.24.2 documentation. (n.d.). Retrieved June 5, 2021, from https://scikit-learn.org/stable/modules/tree.html#mathematical-formulation

Abhishek Sharma. (2020, May 11). Decision Tree vs. Random Forest—Which Algorithm Should you Use? *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

An, G., Xing, M., He, B., Liao, C., Huang, X., Shang, J., & Kang, H. (2020). Using Machine Learning for Estimating Rice Chlorophyll Content from In Situ Hyperspectral Data. *Remote Sensing*, *12*(18), 3104. https://doi.org/10.3390/rs12183104

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

Brownlee, J. (2020, May 24). Recursive Feature Elimination (RFE) for Feature Selection in Python. *Machine Learning Mastery*. https://machinelearningmastery.com/rfe-feature-selection-in-python/

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Hauck, J., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Bakker, D. C. E., Canadell, J. G., Ciais, P., Jackson, R. B., Anthoni, P., Barbero, L., Bastos, A., Bastrikov, V., Becker, M., … Zaehle, S. (2019). Global Carbon Budget 2019. *Earth System Science Data*, *11*(4), 1783–1838. https://doi.org/10.5194/essd-11-1783-2019

Frost, J. (2017, April 2). Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. *Statistics By Jim*. http://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Fu, P., Meacham-Hensold, K., Guan, K., & Bernacchi, C. J. (2019). Hyperspectral Leaf Reflectance as Proxy for Photosynthetic Capacities: An Ensemble Approach Based on Multiple Machine Learning Algorithms. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00730

Garner, R. (2013, July 24). *Seeing Photosynthesis from Space: NASA Scientists Use Satellites to Measure Plant Health* [Text]. NASA. http://www.nasa.gov/content/goddard/seeing-photosynthesis-from-space-nasa-scientists-use-satellites-to-measure-plant-health

Georgios Drakos. (2019, May 23). *Decision Tree Regressor explained in depth*. GDCoder. https://gdcoder.com/decision-tree-regressor-explained-in-depth/

Google Colaboratory. (n.d.). Retrieved June 2, 2021, from https://colab.research.google.com/notebooks/intro.ipynb#recent=true

Guanter, L., Kaufmann, H., Segl, K., Foerster, S., Rogass, C., Chabrillat, S., Kuester, T., Hollstein, A., Rossner, G., Chlebek, C., Straif, C., Fischer, S., Schrader, S., Storch, T., Heiden, U., Mueller, A., Bachmann, M., Mühle, H., Müller, R., … Sang, B. (2015). The EnMAP Spaceborne Imaging Spectroscopy Mission for Earth Observation. *Remote Sensing*, *7*(7), 8830–8857. https://doi.org/10.3390/rs70708830

Gupta, S. (2020, June 23). *Pros and cons of various Classification ML algorithms*. Medium. https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6

Leite, R. (2020, October 7). *Introduction to Cross-Validation: K-Fold*. Medium. https://medium.com/analytics-vidhya/introduction-to-cross-validation-k-fold-7ed9cbd0ed7b

Lutins, E. (2017, September 13). *Grid Searching in Machine Learning: Quick Explanation and Python Implementation*. Medium. https://elutins.medium.com/grid-searching-in-machine-learning-quick-explanation-and-python-implementation-550552200596

Mahanta, J. (2017, July 12). *Introduction to Neural Networks, Advantages and Applications*. Medium. https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-96851bd1a207

Mohammed, G. H., Colombo, R., Middleton, E. M., Rascher, U., van der Tol, C., Nedbal, L., Goulas, Y., Pérez-Priego, O., Damm, A., Meroni, M., Joiner, J., Cogliati, S., Verhoef, W., Malenovský, Z., Gastellu-Etchegorry, J.-P., Miller, J. R., Guanter, L., Moreno, J., Moya, I., … Zarco-Tejada, P. J. (2019). Remote sensing of solar-induced chlorophyll fluorescence

(SIF) in vegetation: 50 years of progress. *Remote Sensing of Environment*, *231*, 111177. https://doi.org/10.1016/j.rse.2019.04.030

Porcar-Castell, A., Tyystjärvi, E., Atherton, J., van der Tol, C., Flexas, J., Pfündel, E. E., Moreno, J., Frankenberg, C., & Berry, J. A. (2014). Linking chlorophyll a fluorescence to photosynthesis for remote sensing applications: Mechanisms and challenges. *Journal of Experimental Botany*, *65*(15), 4065–4095. https://doi.org/10.1093/jxb/eru191

Ray, S. (2017, September 8). Commonly Used Machine Learning Algorithms | Data Science. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Ritchie, H. (2017, October 3). *How much of the world's land would we need in order to feed the global population with the average diet of a given country?* Our World in Data. https://ourworldindata.org/agricultural-land-by-global-diets

Shah, S. H., Angel, Y., Houborg, R., Ali, S., & McCabe, M. F. (2019). A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat. *Remote Sensing*, *11*(8), 920. https://doi.org/10.3390/rs11080920

Slack. (n.d.). *Where work happens*. Slack. Retrieved June 2, 2021, from https://slack.com/intl/en-au/

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225–1231. https://doi.org/10.1016/s0895-4356(96)00002-9

Verrelst, J., Malenovský, Z., Van der Tol, C., Camps-Valls, G., Gastellu-Etchegorry, J.-P., Lewis, P., North, P., & Moreno, J. (2019). Quantifying Vegetation Biophysical Variables from Imaging Spectroscopy Data: A Review on Retrieval Methods. *Surveys in Geophysics*, *40*(3), 589–629. https://doi.org/10.1007/s10712-018-9478-y

Welcome to Python.org. (n.d.). Python.Org. Retrieved June 2, 2021, from https://www.python.org/

Woodgate, W., Suarez, L., van Gorsel, E., Cernusak, L. A., Dempsey, R., Devilla, R., Held, A., Hill, M. J., & Norton, A. J. (2019). tri-PRI: A three band reflectance index tracking dynamic photoprotective mechanisms in a mature eucalypt forest. *Agricultural and Forest Meteorology*, *272–273*, 187–201. https://doi.org/10.1016/j.agrformet.2019.03.020

Woodgate, William, van Gorsel, E., Hughes, D., Suarez, L., Jimenez-Berni, J., & Held, A. (2020). THEMS: An automated thermal and hyperspectral proximal sensing system for canopy reflectance, radiance and temperature. *Plant Methods*, *16*(1), 105. https://doi.org/10.1186/s13007-020-00646-w

Wu, S. (2020, June 14). *What are the best metrics to evaluate your regression model?* Medium. https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b

Zhou, J.-J., Zhang, Y.-H., Han, Z.-M., Liu, X.-Y., Jian, Y.-F., Hu, C.-G., & Dian, Y.-Y. (2021). Hyperspectral sensing of photosynthesis, stomatal conductance, and transpiration for citrus tree under drought condition. *BioRxiv*, 2021.02.26.433135. https://doi.org/10.1101/2021.02.26.433135