A PROJECT REPORT ON

Fine Scale Prediction of Leaf Nitrogen Content and Biomass in Sugarcane

Submitted Towards the
Partial Fulfilment of the Requirements of

Bachelor of Engineering (Computer Engineering)

by

Satyam Kale	Roll No:32
Chaitanya Kasar	Roll No:29
Pushkar Jha	Roll No:30
Rahul Bagate	Roll No:31

Under the Guidance of
Prof. Dr.S.M.Kamalapur

Department of Computer Engineering
K. K. Wagh Institute of Engineering Education & Research
Hirabai Haridas Vidyanagari, Amrutdham, Panchavati, Nashik-422003
Savitribai Phule Pune University

A. Y. 2023-2024 Sem I

K. K. Wagh Institute of Engineering Education and Research
Department of Computer Engineering

CERTIFICATE

This is to certify that the Project Titled

Fine Scale Prediction of Leaf Nitrogen Content and Biomass in Sugarcane

Submitted by

Satyam Kale	Roll No:32
Chaitanya Kasar	Roll No:29
Pushkar Jha	Roll No:30
Rahul Bagate	Roll No:31

is a bonafide work carried out by students under the supervision of *Prof. Dr.S.M.Kamalapur* and it is submitted towards the partial fulfilment of the requirement of Bachelor of Engineering (Computer Engineering) during academic year 2023-2024.

Prof. Dr.S.M.Kamalapur
Internal Guide
Department of Computer Engineering

Prof. Dr. S. S. Sane
Head
Department of Computer Engineering

Abstract

The primary goal of precision agriculture is to improve crop biomass and fertilization management. Crop biomass can be used to determine efficiency of production. Prediction of crop biomass is needed for efficient fertilization management. Estimation of nitrogen content can be used to find nitrogen deficiency in plants. There is a need for the development of a system to predict sugarcane biomass and leaf nitrogen content. UAV platforms and sensing technologies are extensively used in precision agriculture. The work aims to use LiDAR and Multispectral imaging to estimate crop biomass and nitrogen content. Dataset based on UAV LiDAR and multispectral images of sugarcane fields located in Tully river catchment in northeast Queensland will be used for experimentation. The system will predict biomass at harvest using different models like multispectral, LiDAR, and fused predictors, along with a normalized difference vegetation index (NDVI) benchmark and comparing their efficiency and accuracy. This approach holds potential for enhancing precision agriculture practices and optimizing yield management.

Acknowledgment

On behalf of our project team, we would like to extend our heartfelt appreciation to everyone who has played a crucial role in the successful completion of Stage 1 of our project. We are immensely grateful to project guide Prof.Dr.S.M.Kamalapur for her invaluable contributions, guidance, and support during this initial phase. Your expertise and dedication have been instrumental in shaping our project's trajectory. Our team members and colleagues deserve special recognition for their unwavering commitment and hard work. Together, we have tackled challenges, met deadlines, and achieved important milestones that have brought us to this point. We also want to acknowledge the guidance from senior faculty Prof.N.M.Shahane , she has provided us with the support and motivation necessary for our success. This project has been a collaborative effort, and we are thankful for the collective contributions that have made Stage 1 a triumph. As we move forward into subsequent stages, we anticipate your continued support and partnership

Satyam Kale
Chaitanya Kasar
Pushkar Jha
Rahul Bagate
(B.E. Computer Engg.)

Contents

List of Figures

Chapter 1

Introduction

The prime objective of our project is to analyze the given biomass, LIDAR and multispectral data and to build machine learning model for accurate prediction of sugarcane biomass and leaf nitrogen concentration.

1.1 Project Idea

The proposed project aims to advance precision agriculture by developing a system for predicting sugarcane biomass and leaf nitrogen content, crucial factors in optimizing fertilization and overall crop management. Leveraging Unmanned Aerial Vehicle (UAV) platforms equipped with LiDAR and Multispectral imaging technologies. The primary objectives include utilizing LiDAR and Multispectral imaging data to estimate crop biomass and nitrogen content in sugarcane fields. The dataset, derived from UAV observations, serves as the basis for experimentation. The project aims to predict biomass at harvest using various models, including those based on multispectral data, LiDAR data, and a fusion of both, in addition to a normalized difference vegetation index (NDVI) benchmark. Comparisons will be made to assess the efficiency and accuracy of these prediction models. By exploring the strengths of different predictors and their combinations, the project seeks to enhance precision agriculture practices, offering insights into optimal fertilization strategies and overall yield management. The research not only contributes to the scientific understanding of sugarcane growth dynamics but also holds practical implications for sustainable and efficient crop production in the Tully river catchment. Ultimately, the project aims to empower farmers with advanced tools for informed decision-making and improved outcomes in sugarcane cultivation.

1.2 Motivation of the Project

The motivation behind this project stems from the imperative to revolutionize precision agriculture, particularly in the context of sugarcane cultivation. Recognizing the pivotal role of predicting sugarcane biomass and leaf nitrogen content in optimizing fertilization and overall crop management, the project seeks to address existing challenges and propel agricultural practices into a more advanced and efficient era. The utilization of cutting-edge technologies, specifically Unmanned Aerial Vehicle (UAV) platforms equipped with LiDAR and Multispectral imaging, offers an unprecedented opportunity to gather high-resolution data over sugarcane fields in the Tully river catchment in northeast Queensland. This geographical focus reflects the project's commitment to addressing the specific needs and nuances of the local agricultural landscape. The primary motivation is rooted in the necessity for accurate and timely predictions of biomass and nitrogen content, both pivotal factors in the success of sugarcane cultivation. The development of a robust predictive system, drawing insights from LiDAR and Multispectral imaging data, aims to provide farmers with a sophisticated toolset. This toolset will not only enhance their ability to optimize fertilization strategies and overall yield management but also contribute to the broader goals of sustainability and resource efficiency in agriculture. By comparing and evaluating various predictive models, including those based on multispectral data, LiDAR data, and their fusion, the project aspires to unravel the strengths and efficiencies inherent in each approach. This exploration holds promise for refining precision agriculture practices, guiding farmers towards informed decision-making, and ultimately leading to improved outcomes in sugarcane cultivation. The motivation extends beyond scientific inquiry; it embraces the pragmatic realm of empowering farmers. The research seeks to translate theoretical advancements into practical applications, offering tangible benefits for those engaged in sugarcane production. Ultimately, the project aspires to be a catalyst for positive change in the agricultural landscape, marrying technological innovation with on-the-ground effectiveness for sustainable and efficient crop production.

1.3 Literature Survey

Given below are the projects we have studied and thus they increase the clearity regarding our project understanding and decision making.

Rahman et. al., 2013 explored the capacity of a random forest (RF) regression technique to choose the spectral characteristics from hyperspectral data that are required to forecast the N content in sugarcane leaves. Two Hyperion photos were taken from fields of sugarcane that was 6-7 months old, variety N19, in order to accomplish this. The spectral data was analysed using the machine-learning RF algorithm as a feature-selection and regression technique. The content of sugarcane leaf N was also predicted using stepwise multiple linear (SML) regression after the redundancy in the hyperspectral data was reduced. Both non-linear RF regression (coefficient of determination, R 2 = 0.67; root mean square error of validation (RMSEV) = 0.15) and SML regression models (R2 = 0.71; RMSEV = 0.19), according to the results, can be used to predict the N content in sugarcane leaves.

Andrews et.al., 2013 explained the effects of nitrogen (N) form on plants at various developmental stages. Specifically, nitrate (NO3-) was contrasted with the other types of N that plants utilise. It is determined that plants' root architecture, dry matter partitioning between shoot and root, leaf growth and function, and seed germination time and rate can all be impacted by the type of N that is accessible to them. The extent to which these effects occur depends on environmental variables other than the availability of N. These effects have a varied mechanism. More research is needed to determine the significance of root or shoot NO3-assimilation in various environmental settings.

Ballester et. al., 2017 assessed the effectiveness of spectral indices derived from a drone-based system for monitoring nitrogen levels and predicting cotton yield on a commercial farm. Various fertilization methods and spectral data were used throughout the growing season. The study found that certain indices, such as SCCCI and NDRE, were useful for tracking nitrogen levels and predicting yield, with their effectiveness varying at different stages of the crop's development. Overall, research highlights the potential of using unmanned aerial systems for real-time nitrogen monitoring in cotton farming, while also recognizing the challenges in making early-season fertilization recommendations based on spectral data.

The study conducted by Bendig et. al., 2014 aimed to explore the use of super-high-resolution aerial imagery obtained from a small Unmanned Aerial Vehicle (UAV) for estimating fresh and dry biomass in barley crops with varying cultivars and nitrogen treatments. The UAV im-

agery was used to measure plant height, an essential indicator of crop growth, across the different barley cultivars and nitrogen treatment plots. The study compared the plant height measurements obtained from UAV imagery with reference measurements, such as ground-based measurements or traditional methods. The strong correlation with an R-squared value of 0.92 indicated that UAV-based measurements were highly accurate. The research used the plant height data to estimate fresh and dry biomass for the 18 barley cultivars under two different nitrogen treatments. The resulting R-squared values of 0.81 for fresh biomass and 0.82 for dry biomass suggested that the UAV imagery had significant potential for biomass estimation.

Maresma et. al., 2016 analyzed crop height (derived from a multispectral compact camera mounted on a UAV), crop height based on green-band and red-band based multispectral vegetation indices, and SPAD to predict grain yield. Maize experimental plots were subjected to seven distinct inorganic N rates (0, 100, 150, 200, 250, 300, and 400 kg·N·ha1), two distinct pig slurry manure rates (Ps) (150 or 250 kg·N·ha1), and four distinct inorganic-organic N combinations (N100Ps150, N100Ps250, N200Ps150, and N200Ps250). The spectral index whose ultimate grain yield was most adequately explained for the Wide Dynamic Range Vegetation Index (WDRVI) was the N treatments. A critical threshold above/below 250–300 kg·N·ha1 was found. The WDRVI, NDVI, and crop height did not exhibit any noteworthy changes in response to further N application at the economically optimal fertilisation rate of 239.8 kg·N·ha1, which resulted in a grain yield of 16.12 Mg·ha1. This indicates that they have the ability to anticipate yields at the V12 stage. In order to reduce ambiguity, a ranking of various vegetation indices and crop height is finally suggested. This eliminates the need to base judgements just on one measure.

According to Sanches et. al., 2018 RGB images captured by unmanned aerial vehicles have the potential to be used for yield assessment and prediction in sugarcane fields. The study addresses the challenge of estimating crop yield, particularly in sugarcane fields, by harnessing the capabilities of Unmanned Aerial Vehicles (UAVs). The research focuses on assessing canopy closure in sugarcane using UAV imagery, comparing different planting approaches and row-spacing treatments. Key growth stages were evaluated, and two indices, Leaf Area Index (LAI) and Green-Red Vegetation Index (GRVI), were employed, with GRVI from UAVs proving more effective in reflecting crop yield con-

ditions due to its superior spatial resolution (R2 = 0.69, compared to LAI's R2 = 0.34). Combining these indices improved yield estimates by 10% (R2 = 0.79), demonstrating the cost-effective and high-precision utility of UAV imagery in aiding growers' decision-making for sugarcane cultivation.

Zhou et. al., 2017 focused on the timely and non-destructive assessment of rice grain yield through the utilization of remote sensing, particularly leveraging the capabilities of unmanned aerial vehicles (UAVs) to gather high-resolution imagery over a regional scale. The research investigates the prediction of rice grain yield using vegetation indices (VIs) derived from both multispectral (MS) and digital images, emphasizing the temporal aspect by considering single-stage and multi-temporal VIs. The study identifies the booting stage as the most suitable phase for grain yield prediction using VIs from both digital and MS images. This stage is crucial in rice development and serves as a strong indicator of yield potential. The research compares various VIs to assess their effectiveness in predicting grain yield. For single-stage predictions, the Vegetation Atmospherically Resistant Index (VARI) from digital images and the Normalized Difference Vegetation Index (NDVI[800,720]) from MS images are singled out as the most promising candidates, with R-squared (R2) values of 0.71 and 0.75, respectively. This indicates the strong correlation between these VIs and grain yield. The study also explores the use of multi-temporal VIs, which consider data from multiple growth stages. These multi-temporal VIs show a higher correlation with grain yield compared to single-stage VIs. Among them, the Multiple Linear Regression function (MLR) that combines VIs at two random growth stages performs the best, with the highest correlation coefficients of 0.76 with MLR(NDVI[800,720]) and 0.73 with MLR(VARI). The study further demonstrates that VIs with a high correlation with Leaf Area Index (LAI), a measure of plant canopy density, are effective in predicting yield. Additionally, VIs composed of bands in the red edge (720 nm) and near-infrared (800 nm) wavelengths prove to be particularly useful in estimating both grain yield and LAI, especially at a high level.

Chapter 2

Problem Definition and scope

The project is precisely formulated to achieve accurate predictions of biomass and leaf nitrogen concentration. The scope is specifically confined to outlining the methods and processes essential for constructing a highly effective prediction system.

2.1 Problem Statement

The project aims to address the critical need in precision agriculture for predicting sugarcane biomass and leaf nitrogen content using UAV LiDAR and Multispectral imaging. The goal is to develop a system that efficiently utilizes these technologies. The project seeks to employ various models, including multispectral, LiDAR, and fused predictors, along with a normalized difference vegetation index (NDVI) benchmark, to predict sugarcane biomass at harvest. By comparing their efficiency and accuracy, the research endeavors to enhance precision agriculture practices, particularly in optimizing yield management for sustainable sugarcane production.

2.1.1 Goals and objectives

Given below are the primary aims and targeted procedures.

1. Study LiDAR and Multispectral Data: The first objective is to thoroughly examine the data collected through LiDAR (Light Detection and Ranging) and multispectral imagery. This initial step involves data acquisition and data quality assessment.

2. Data Cleaning and Feature Selection: After gathering the data, the next step is to identify and rectify any outliers or errors within the dataset. Outliers are data points that significantly deviate from the norm and may distort analysis. Error correction and data cleaning are essential to ensure the accuracy and reliability of subsequent analyses. Additionally, the project aims to include only the most relevant attributes and features from the dataset. This step involves data preprocessing and feature selection techniques.

3. Develop Predictive Model: The core goal of this project is to create a predictive model. This model will use the cleaned and selected data to make predictions about the stage at which biomass prediction and leaf nitrogen content can be derived. Developing this model involves selecting an appropriate machine learning algorithm, training it with the data, and fine-tuning it to achieve accurate predictions. This could be done using Ordinary Least Squares (OLS) methods.

4. Model Evaluation with F-Test and Benchmarks: To assess the performance and reliability of the predictive model, various evaluation methods will be employed. One of them is the F-test, a statistical test used to compare the variances of different groups within the data. This can help determine the significance of different variables in predicting the outcomes. Additionally, the project will compare the predictive model's performance against various benchmarks like avg NDVI and max NDVI, which might include other data like SRA data.

5. Deployment of the Predictive Model: The final objective is to put the developed predictive model into practical use. Deployment involves integrating the model into a usable and accessible system, such as software or a web application, where it can provide real-time predictions or analysis based on new data. This deployment phase aims to make the insights and predictions derived from the model readily available and useful for relevant stakeholders, such as farmers or researchers.

2.1.2 Scope

The extent to range and extent of our project is limited to the given points :-

1)To create the model that will predict:
   a) Leaf Nitrogen content
   b) Sugarcane Biomass

2)To test the ability of multispectral imagery in predicting leaf nitrogen concentrations

3)To investigate benefits of fusing LiDAR and Multispectral data for predicting sugarcane biomass

2.2 Methodology

Our methods are primarily focused on performing the data operations and building of accurate predictive models.

1)Experimental design:
To create the model that will predict:
   Leaf Nitrogen content
   Sugarcane Biomass

2) Field measurements: Using multispectral imagery 10 vegetation indices were calculated which further evaluated in predicting sugarcane leaf N content and Biomass. For LiDAR scans after preprocessing steps ,is with reference to ground surface and used to extract 48 predictor variables(i.e Features). for Multispectral imagery 70 predictors variables(Features) were calculated i.e 7 features for 10 vegetation indices (7X10). Therefore the total predictor variables (features) was 48 of LiDAR + 70 of Multispectral = 118.

3) Dimension Reduction: The final data set for the total biomass dependent variable contained n=56 observations and p=118 predictors in total (i.e 70 multispectral predictors(features) and 48 LiDAR predictors in each of six surveys). Since no. of predictor variables(P)was less than no.of observations(N) i.e P much less than N. As the sample size of n=56 as relatively small , methods that required a training / test split for biomass prediction were avoided, as further reduction in sample size would affect the model's ability to fit the data well, and test error would be sensitive to the initial choice for the split. . Therefore, principal components analysis (PCA) should be used to reduce the dimension of the information reduced to a few predictors for each of the

Predictor variables extracted from normalized (i.e. with respect to ground surface) LiDAR-derived point clouds.

Feature	Explanation
max	Maximum height
avg	Average height
qav	Average square height
std	Standard deviation of height
ske	Height skewness
kur	Height kurtosis
p05 to p95	5th to 95th height percentiles (increments of 5 percentiles)
b05 to b95	5th to 95th bincentiles[a] (increments of 5 bincentiles)
d00[b]	The number of points between 0 (i.e. ground) and 0.01 m divided by the total number of points
d01[b]	The number of points between 0.01 and 0.5 m divided by the total number of points
d02[b]	The number of points between 0.5 and 1 m divided by the total number of points
d03[b]	The number of points between 1 and 10 m divided by the total number of points

[a] Fraction of points between ground and the height percentile.

[b] Threshold values for d00, d01, d02 and d03 were defined to represent penetration of laser pulses at different height levels of sugarcane.

Figure 2.1: LiDAR Predictor Variables

Vegetation indices calculated from various Micasense RedEdge reflectance band combinations.

Name	Equation
Normalized Difference Vegetation Index (NDVI)	$(NIR - R)/(NIR + R)$
Normalized Difference Red Edge Index (NDRE)	$(NIR - RE)/(NIR + RE)$
Green NDVI (GNDVI)	$(NIR - G)/(NIR + G)$
Enhanced Vegetation Index (EVI)	$2.5(NIR - R)/(NIR + 6R - 7.5B + 1)$
Modified Anthocyanin Content Index (MACI)	NIR/G
Optimized Soil Adjusted Vegetation Index (OSAVI)	$(1 + 0.16)(NIR - R)/(NIR + R + 0.16)$
Simplified Canopy Chlorophyll Content Index (SCCCI)	$NDRE/NDVI$
Transformed Chlorophyll Absorption and Reflectance Index (TCARI)	$3[RE - R - 0.2(RE/G)(RE/R)]/OSAVI$
Triangular Greenness Index (TGI)	$-0.5[(668 - 475)(R - G) - (668 - 560)(R - B)]$
Visible Atmospherically Resistant Index (VARI)	$(G*R)/(G + R - B)$

R = red, G = green, B = blue, RE = red edge, NIR = near infrared.

Figure 2.2: Vegetation Indices

multispectral and LiDAR data sets respectively. PCA constructs a set of linear combinations of all the predictors (i.e. principal components). The weights in each linear combination are chosen such that the principal component has the largest variance and is orthogonal to all previous principal components. By retaining only the first few principal components as predictors, the dimension of the data was greatly reduced. PCA was undertaken on the multispectral and LiDAR data separately. For each data set, observations were grouped across time, resulting in 336 observations (n=56 biomass samplings * 6 surveys conducted = 336 observations) for each of 70 multispectral predictors(Features), and 336 observations for 48 LiDAR predictors(Features). To choose the number of components based only on the predictors and not the dependent variable, this corresponded to four and three principal components in the multispectral and LiDAR data sets, respectively.

4) Biomass Prediction: In this study, the dependent variable is at-harvest biomass, comprising 56 observations. Prediction will involve four principal components from multispectral data and three from LiDAR data across six surveys, totaling 42 predictors (seven principal components across six time steps), aiming to model and forecast biomass based on these input features. Instead, our objective is to predict the at-harvest biomass independently in each of the six Surveys. For this a set of five different linear regression models in each of the six Surveys will be estimated. Models 1 and 2 serve as the benchmark, containing only NDVI (avg and max, respectively). Models 3 and 4 will use only the multispectral or LiDAR principal components, respectively. Model 5 will be the multispectral/LiDAR fusion. In addition, each model will be restimated including a term for the N application rate to determine whether the additional knowledge of the N fertilizer application rate improves predictive power. Model performance will be evaluated using the adjusted coefficient of determination $\text{Ra}\hat{2}$. Total, leaf and stem biomass as well as number of stems were measured independently in the biomass dataset.

5) Leaf N prediction: In the assessment of leaf N content as a dependent variable n = 160 observations were utilized (i.e. 40 observations in four time periods (i.e. Surveys 2–5)) as mentioned in dataset section. As LiDAR derived information was not expected to be informative of leaf N content, we did not investigate LiDAR predictors and investigated only multispectral imagery derived vegetation indices of the leaves sampled in each Survey. While the set of 70 predictor

Predictor variables extracted from multispectral imagery derived vegetation indices.

Statistics	Explanation
max	Maximum value of all pixels within each plot
min	Minimum value of all pixels within each plot
avg	Average value of all pixels within each plot
std	Standard deviation of all pixels within each plot
p25	25th percentile of all pixels within each plot
p50	50th percentile of all pixels within each plot
p75	75th percentile of all pixels within each plot

Figure 2.3: Multispectral predictor variables

Table 1
Survey dates and field samples collected.

Event	Survey date	Survey interval (days)	Days after harvest (DAH)	Leaf *N* sampled	Biomass sampled	Phenological phase
Survey #1	8/11/2017	–	58	No	No	Early growth
Survey #2	20/12/2017	42	100	Yes	No	Early growth
Survey #3	31/01/2018	42	142	Yes	No	Maturation
Survey #4	15/03/2018	43	185	Yes	No	Maturation
Survey #5	26/04/2018	42	227	Yes	No	Maturation
Survey #6	6/06/2018	41	268	No	Yes	Flowering

Note: Survey #1 was performed right after sugarcane fertilization and at the outset of the early growth phase, while Survey #3 was performed at the transition of early growth and maturation phenological phases.

Figure 2.4: Table of survey data and field samples

variables constructed from the multispectral imagery derived vegetation indices were identical to those constructed for biomass prediction, the sampling plots used to generate predictor variables were different . Since in each Survey there were 70 predictors for 40 observations of the dependent variable, a dimension reduction using PCA should also be performed. For each observation in the leaf N data set, the set of 70 predictors was projected through the principal components model estimated previously for multispectral predictors in biomass prediction. This resulted in the conversion of the 70 predictors into four principal components for each observation. Next, a linear regression will be utilized in which leaf N content will be regressed against each of the four contemporaneous principal components and a quadratic time trend.

2.3 Outcome

The system should be able to accurately predict sugarcane biomass and estimate leaf nitrogen content. Our goal is to independently forecast the biomass at harvest for each of the six surveys. Each of the six Surveys will have a set of five distinct linear regression models estimated for this purpose.The benchmark models, 1 and 2, only have the NDVI (avg and max, respectively). Only the multispectral or LiDAR primary components, respectively, will be used in Models 3 and 4. The multispectral/LiDAR fusion will be Model 5. Furthermore, to ascertain whether the new knowledge of the N fertiliser application rate increases predictive power, each model will be restimulated with a term for the N application rate. The adjusted coefficient of determination $\text{Ra}\hat{2}$ will be used to assess the performance of the model. The leaf N content will then be regressed against each of the four contemporaneous principal components and a quadratic temporal trend using a linear regression.

2.4 Type of Project

Our project falls within the realm of a research project, and its domains are specifically confined to the application and scope defined by the project itself.

Research Project - The type of project described is a research project in the domain of precision agriculture, specifically focusing on sugarcane cultivation. The project falls within the broader category of agricultural technology and remote sensing applications for crop management.

Domain - This project is situated in the domain of precision agriculture, specifically utilizing machine learning methodologies. It encompasses the application of advanced algorithms, including Principal Component Analysis (PCA) for dimensionality reduction, along with rigorous data preprocessing techniques. The focus lies on regression methods for predicting sugarcane biomass and estimating leaf nitrogen content. Additionally, the project involves thorough performance testing of the generated machine learning models to ensure their accuracy and efficiency in optimizing yield management practices within the agricultural context.

Chapter 3

Project Plan

Our project plan is simple yet efficient. A brief overview of project timeline and team organization is mentioned below :-

3.1 Project Timeline

A project timeline chart, or Gantt chart, is a visual tool used in project management to display project tasks and their timing. It shows tasks as bars over time, with their start and end dates, dependencies, milestones, and progress. It helps plan, track, and communicate project schedules.

3.2 Team Organization

The team consist of 3 distinct elements namely, the project mentor, the project leader, and the team members.
1. Project Mentor Prof. Dr.S.M.Kamalapur is our mentor. She helps in providing guidance over project guidance and implementation. She also reduced the errors and proliferates the workflow.
2. Team leader Rahul Bagate is our team leader. He motivates the whole team for building an ideal project. He is the final decision-maker.
3. Team member(s) The ground implementation is carried out by them. Here, Chaitanya Kasar, Pushkar Jha, and Satyam Kale forms part of the team. Pushkar Jha - He main work is to analyse the data and reduce the errors and redundancy in the project. A prime member for developing the diiferent machine learning model. Satyam Kale - Designing the algorithm, the solution's flow is his key role. Chaitanya Kasar - His critical thinking skills aid in projects proliferation towards acheiving perfection. He is the significant member for articulating the documents.

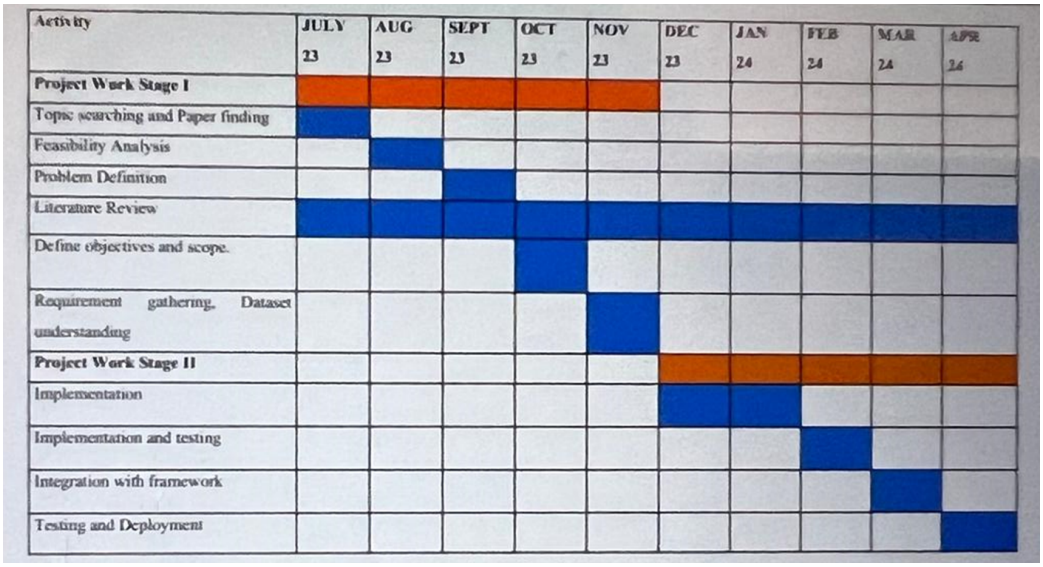

Figure 3.1: Project Timeline

Chapter 4

Software requirement specification

Software requirement specification outlining software functionality, constraints, and user expectations.

4.1 Functional Requirements

The functional requirement(s) describes a particular behavior or functionality of the system. Or, in other words, they describes what the system should do.
1) The system should integrate LiDAR and multispectral data to create a comprehensive dataset for analysis.
2)The system should provide a method for estimating sugarcane biomass using LiDAR and multispectral data.
3)The system should offer a mechanism for estimating leaf nitrogen content in sugarcane plants based on spectral data

4.2 Non Functional Requirements

The dependability and sustainability for efficient project building are mentioned below :-

1)Availability:
The Leaf Nitrogen and Biomass CSV with the multispectral and LiDAR Predictor variables are available for the system.
2)Reliability:

The system must consistently provide reliable and accurate estimations of Leaf Nitrogen Content and Total biomass.
3)Maintainability:
Ensuring clear documentation and modular code structure for ease of updates. Regularly validating and adapting models with new data for sustained accuracy.
4)Security:
Implement strong encryption protocols for data both in transit and at rest to safeguard sensitive sugarcane-related information.

4.3 Assumptions

The ensuing points encapsulate our underlying assumptions:
1)The UAV LiDAR and multispectral imagery data are assumed to be accurate and representative of the actual conditions of the sugarcane fields.
2)Compatible and sufficient hardware and software resources available.

4.4 Requirements

Here are some key considerations for requirements from various viewpoints:

4.4.1 Hardware Requirements

Given below are the hardware requirements from the developer's viewpoints :-

- Advanced processing devices, whether in the form of laptops or computers, can facilitate the seamless execution of the project.
- Every developer should possess a computing device with a minimum storage capacity of 256 GB and a RAM of at least 8 GB.

4.4.2 Software Requirements

Given below are the software requirements from the developer's viewpoints:-

- Essential software tools for tasks such as data processing and machine learning include applications like Jupyter Notebook or Google Colab, Python 3, and Microsoft Excel software.

- Collaborative Platforms: Platforms that facilitate collaboration among developers and stakeholders, promoting efficient communication and information sharing like github.

Chapter 5

Detailed Design

The figures and information in this section has a detailed explanation of the project's design and planning.

5.1 Architectural Design(Block Diagram)

The Figure 5.1 and 5.2 given below shows a diagram of Leaf Nitrogen prediction and Biomass prediction with the following components:
• Pre-processing: The input data of Multispectral and lidar samples is pre-processed to remove noise, redundancy and required variables.
• Principal Component Analysis, or PCA It is necessary because our data's sample size is so small. That's why it will be the best practice to perform PCA.
• Model Building: In this component One key predictor variable is chosen and model is trained and test accordingly.that's there are three prediction models are there.(Max NDVI, Min NDVI, Avg NDVI).

5.1.1 Biomass Prediction

The biomass data description is explained such that it gives a clear idea of the CSV format files.

Biomass Data: Number of biomass sampling plots:28 The total of n = 56 (28X2) observations of at-harvest biomass were used as a dependent variable. total biomass, leaf biomass and stem biomass are being predicted.
total – surveys samples fields
6 X 28 X 2 = 336

The analysis involves assessing leaf nitrogen (N) content as the dependent variable with 160 observations collected during four time periods (Surveys 2–5), each containing 40 observations. The goal is to understand how multispectral imagery-derived vegetation indices relate to leaf N content.
Here are the key steps of the analysis:
1. Selection of Predictors: - Multispectral imagery and LiDAR imagery-derived vegetation indices are chosen as potential predictor variables for biomass content prediction.
2. PCA Dimension Reduction: - Although there are 70 predictor variables derived from multispectral imagery and 48 predictor variables from LiDAR imagery, a dimension reduction technique, Principal Component Analysis (PCA), is applied. The purpose of PCA is to transform the original set of variables into a smaller set of uncorrelated variables called principal components.
3. Conversion to Principal Components: - For each observation in the biomass dataset, the set of 118 predictor variables is projected onto the principal components generated by the PCA model. This transformation results in a reduced set of predictor variables, where the 118 variables are replaced by a smaller number of principal components. - In this case, seven principal components are derived for each observation. These principal components represent a compressed and more informative representation of the original multispectral and LiDAR data.
4. Linear Regression: - With the reduced set of predictor variables (seven principal components) and an additional quadratic time trend (indicating changes over time), a linear regression model is constructed.
The structure of the predictive models is as follows:
1. Benchmark Models (Models 1 and 2): - Model 1 will include only the average Normalized Difference Vegetation Index (NDVI), while Model 2 will include only the maximum NDVI. These models serve as the baseline for comparison.
2. Predictive Models (Models 3, 4, and 5): - Model 3 will solely utilize the four principal components derived from multispectral data, while Model 4 will use the three principal components from LiDAR data. - Model 5 is a fusion model, combining the information from both multispectral and LiDAR principal components.
3. Incorporating N Fertilizer Application: - Each of these models will be re-estimated with an additional term for the N fertilizer application rate. This step aims to determine whether incorporating information about N application enhances the models' predictive capacity.
4. Model Performance Evaluation: - Model performance will be assessed using the adjusted coefficient of determination (Ra^2). This metric accounts for

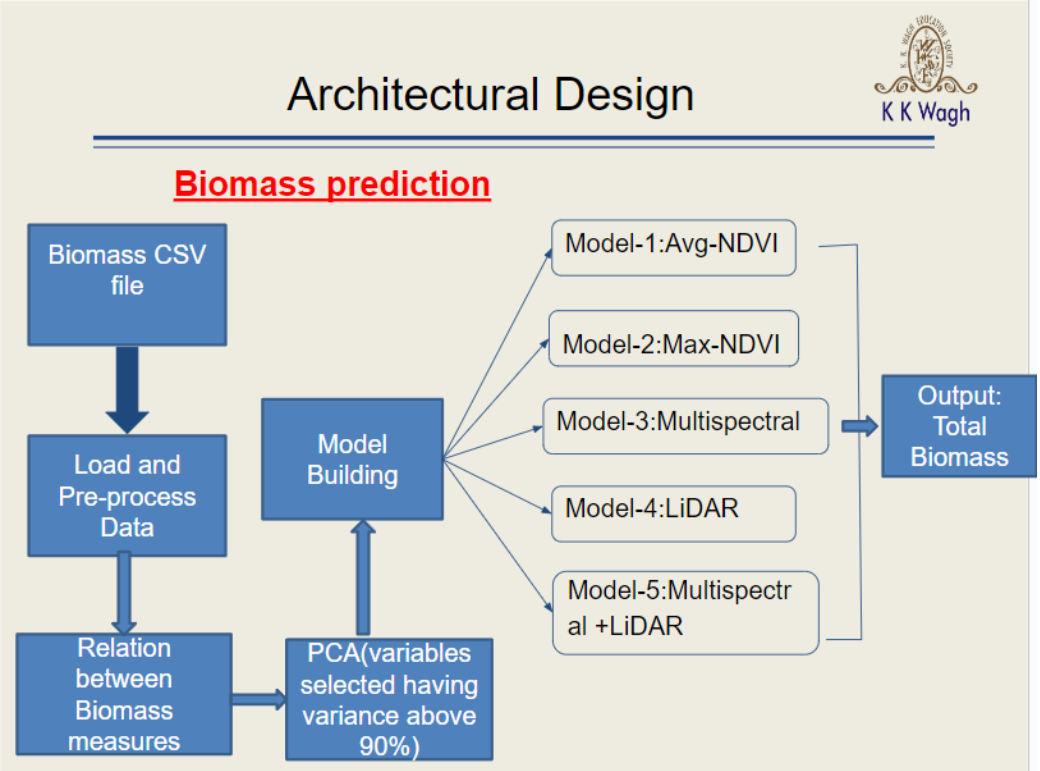

Figure 5.1: Biomass Prediction

the number of predictors in the model and provides a more reliable measure of the model's goodness of fit.
Regarding the biomass dataset, it includes measurements of total biomass, leaf biomass, stem biomass, and the number of stems.

5.1.2 Leaf Nitrogen Prediction

The leaf nitrogen data description is explained such that it gives a clear idea of the CSV format files.

Leaf-N-Data: Number of leaf N sampling plots:20 In the assessment of leaf N content as a dependent variable n = 160 observations were utilized (i.e. 40 observations in four time periods (i.e. Surveys 2–5)).
70 predictor variables constructed from the multispectral imagery derived vegetation indices
The analysis involves assessing leaf nitrogen (N) content as the dependent variable with 160 observations collected during four time periods (Surveys

2–5), each containing 40 observations. The goal is to understand how multispectral imagery-derived vegetation indices relate to leaf N content.
Here are the key steps of the analysis:
1. Selection of Predictors: - Multispectral imagery-derived vegetation indices are chosen as potential predictor variables for leaf N content prediction.
2. PCA Dimension Reduction: - Although there are 70 predictor variables derived from multispectral imagery, a dimension reduction technique, Principal Component Analysis (PCA), is applied. The purpose of PCA is to transform the original set of variables into a smaller set of uncorrelated variables called principal components. - The PCA model was estimated previously for biomass prediction. However, it's now applied to leaf N content prediction. This model determines how the 70 predictor variables can be combined to create new, orthogonal variables (principal components) that capture the most significant variation in the data.
3. Conversion to Principal Components: - For each observation in the leaf N dataset, the set of 70 predictor variables is projected onto the principal components generated by the PCA model. This transformation results in a reduced set of predictor variables, where the 70 variables are replaced by a smaller number of principal components. In this case, four principal components are derived for each observation. These principal components represent a compressed and more informative representation of the original multispectral data.
4. Linear Regression: - With the reduced set of predictor variables (four principal components) and an additional quadratic time trend (indicating changes over time), a linear regression model is constructed. - Leaf N content is regressed against these predictors to establish a relationship between the principal components, the quadratic time trend, and leaf N content. This model aims to quantify how leaf N content is influenced by these factors.

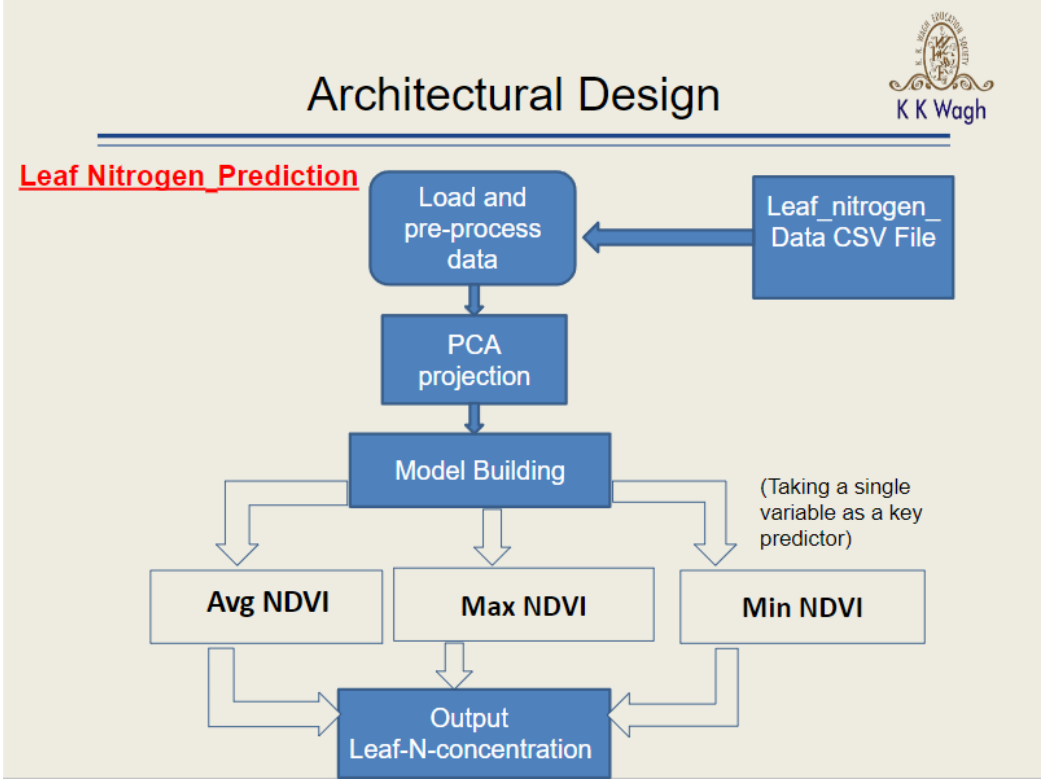

Figure 5.2: Leaf Nitrogen Prediction

Chapter 6

Experimental setup

Experimental setup for the project will require following technologies

6.1 Technologies used

The software and tools required to implement our project are :-

6.1.1 Jupyter Notebook

The project will use Jupyter Notebooks to preprocess and clean the LiDAR and multispectral data, perform exploratory data analysis, develop and test machine learning models, and document the code and findings.

6.1.2 Google Colaboratory (Colab)

Colab allows access to powerful GPUs and TPUs for machine learning tasks, enabling you to train complex models efficiently, which is crucial for tasks like leaf nitrogen content and biomass prediction.

6.1.3 Python / Python 3

Python is the primary programming language for data science and machine learning projects. This project will use Python to implement data preprocessing, feature engineering, model development, and evaluation. Popular Python libraries like NumPy, pandas, scikit-learn, and TensorFlow/PyTorch (for deep learning) can be leveraged to perform these tasks effectively.

6.2 Data Set

Five N treatments (0, 70, 110, 150, 190 kg N/ha) were replicated four times using a randomized complete block design.Each block was 10 m wide, 30 m long and consisted of six sugarcane rows. Six UAV surveys were conducted every six weeks, starting in November 2017 and ending in June 2018, to track sugarcane growth changes from ratooning to harvest over time.
Datasets:

6.2.1 Leaf-N-Data

In the assessment of leaf N content as a dependent variable.Leaf N was measured during Surveys 2–5.Leaf sampling involved selecting leaves from stems of average height, sampling the third leaf from the top of the stem and collecting 40 leaves at random from across rows 2 and 5 of each N application block. The third leaf from the top of the stem was selected for further analysis. n = 160 observations were utilized (i.e. 40 observations in four time periods (i.e. Surveys 2–5)). 70 predictor variables constructed from the multispectral imagery derived vegetation indices. (i.e. 10 multipspectral vegetation indices * 7 Predictor variables = 70).

6.2.2 Biomass-Data

The total of n = 56 observations of at-harvest biomass were used as a dependent variable. Total biomass, leaf biomass and stem biomass are being predicted. Biomass sampling was performed in 56 randomly distributed 2 m $\times$ 2 m plots along rows 1 and 6 immediately following Survey 6 (i.e. 7 weeks before the sugarcane was due to be harvested). Total, leaf, and stem fresh biomass (kg) as well as number of stems were measured in each 2 m $\times$ 2 m plot. The final data set for the total biomass dependent variable contained n = 56 observations and p = 118 predictors in total (i.e.70 multispectral predictors and 48 LiDAR predictors in each of six Surveys). Total of 336(56x6) observations were obtained in all Six Surveys that were conducted.

6.3 Performance Parameters

The projects durability and productivity are based on the given factors:-

6.3.1 Accuracy

The precision and the methods to proliferate the projects accuracy are discussed :-

a. Adjusted R-squared:
Adjusted R-squared : 1 - (1 - r2)*((n - 1) / (n - p))

6.3.2 Precision

The precision formula is given by:-
Precision = True positives/ (True positives + False positives)

6.4 Efficiency Issues

The issues regarding our project mainly revolve around splitting of the data for training and testing purposes, the low sample size of the data, data integration and statistical inefficiency.

1. Sample Size Limitations: A sample size of 56 is relatively small, which can impact the robustness of statistical models. When working with a small sample, there's an increased risk that the results might not accurately represent the entire population. Statistical models, such as regression or machine learning algorithms, may struggle to identify meaningful patterns or relationships in the data. This is because the sample might not capture the true variability and complexity of the population, and any noise or outliers in the data can have a disproportionate impact. Consequently, the models may not perform well in terms of predictive accuracy.

2. Training/Test Split Concerns: In typical machine learning workflows, a dataset is divided into a training set and a test set to evaluate a model's generalization performance. However, with a small sample size like 56, there's a practical concern about splitting the data, as it further reduces the amount of data available for training. This can lead to overfitting, where the model essentially memorizes the training data but doesn't generalize well to new, unseen data. On the other hand, not using a test set can make it challenging to estimate how well the model will perform on new data or in real-world situations, as there is no independent dataset for evaluation.

3. Fusion of Data Sources and Synchronization: Combining data from multiple sources, such as LiDAR and multispectral data, can be complex. Ensuring that the data seamlessly integrates, is accurate, and aligns properly is critical. Misaligned or mismatched data can introduce errors in predictions. For example, if LiDAR data and multispectral data are not synchronized correctly, it can result in incorrect spatial relationships, making it challenging to extract meaningful information or create accurate models. Data fusion techniques and precise calibration are essential to avoid these errors.

4. Statistical Inefficiency (Relation between n and p): When "p" (the number of predictors or features) is approximately equal to or greater than "n" (the number of samples), it leads to a statistical inefficiency problem. In this scenario, there are more variables to estimate than there are data points to estimate them, which can cause instability and unreliable parameter estimates in statistical models. It can also make it difficult to identify which predictors are truly important, leading to high uncertainty in model results. Regularization techniques and feature selection are often used to address this issue by reducing the number of predictors and improving the model's stability and generalization.

Chapter 7

Summary and Conclusion

The summary of our projects serves the prime objective to showcase that how our project can be more accurate than the previous methods for prediction of sugarcane biomass and leaf nitrogen. The conclusion gives an overview of our work done till now and some insights about our future planning.

7.1 Summary

Predicting leaf nitrogen and biomass using hyperspectral and lidar data offers advantages over the study by Abdel Rahman et al. Their research employed random forest (RF) regression to forecast leaf nitrogen content in sugarcane leaves based on hyperspectral data, which achieved an R2 of 0.67. However, using multispectral and lidar imagery can provide more comprehensive information, enabling accurate leaf nitrogen and biomass predictions. Utilizing multispectral and lidar imagery for leaf nitrogen and biomass prediction offers several advantages over the study by Andrews, Raven, and Lea. While their research focused on the effects of different nitrogen forms on plant development, the use of advanced imagery technologies. The studies that were performed earlier also didn't use as many vegetation indices that we have obtained from multispectral imagery. This system also has an additional advantage of variation of data (i.e. The data that is provided from the six surveys that have been performed for prediction). The system will perform Principal Component Analysis (PCA) on two datasets: multispectral and LiDAR data. PCA is used to reduce data dimensionality and extract essential components. This will be done by visualizing the explained variance for these components to determine how many to retain. It will create plots showing explained variance against the number of components for both datasets and will calculates the cumulative explained variance.

7.2 Conclusion

After a comprehensive study and analysis of the entire dataset, we have formulated a conclusive plan outlining the flow and tasks ahead. Our primary objective in this study is to showcase the potential of UAV LiDAR and multispectral imagery in predicting sugarcane biomass and leaf nitrogen (N) content. We are in the process of constructing a model with the goal of achieving a higher R2 value. Multiple models for biomass prediction are being developed, and the one demonstrating the highest accuracy score will be implemented. For predicting Leaf Nitrogen concentrations, we will utilize multispectral data in CSV format as input, constructing a regression model for accurate prediction.

References

- Abdel-Rahman, E.M., Ahmed, F.B., Ismail, R., 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. Int. J. Rem. Sens. 34 (2), 712–728.

- Andrews, M., Raven, J., Lea, P., 2013. Do plants need nitrate? The mechanisms by which nitrogen form affects plants. Ann. Appl. Biol. 163 (2), 174–199.

- Whitehead, K., Hugenholtz, C.H., 2014. Remote sensing of the environment with small unmanned aircraft systems (UASs). Part 1. A review of progress and challenges. J.Unmanned Vehicle Syst. 2 (3), 69–85.

- Bendig, J., Bolten, A., Bennertz, S., Broscheit, J., Eichfuss, S., Bareth, G., 2014. Estimating biomass of barley using crop surface models (CSMs) derived from UAV-based RGB imaging. Rem. Sens. 6 (11), 10395–10412.

- Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., Sousa, J., 2017. Hyperspectral imaging: a review on UAV-based sensors, data processing and applications for agriculture and forestry. Rem. Sens. 9 (11), 1110.

- Ballester, C., Hornbuckle, J., Brinkhoff, J., Smith, J., Quayle, W., 2017. Assessment of inseason cotton nitrogen status and lint yield prediction from unmanned aerial system imagery. Rem. Sens. 9 (11), 1149.

- Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput. Electron. Agric. 151, 61–69.