MULTI-SENSOR FUSION FOR TROPICAL FOREST CARBON STOCK PREDICTION

Conduct by:

Name : Vladimirrahman Salsabil Abdullah
Student ID : 22523298

PROGRAM STUDI INFORMATIKA – PROGRAM SARJANA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ISLAM INDONESIA
2026

SUPERVISOR ENDORSEMENT PAGE

Multi-Sensor Fusion for Tropical Forest Carbon Stock Prediction

THESIS

Conduct By:

Name : Vladimirrahman Salsabil Abdullah
Student ID : 22523298

Yogyakarta, January 07, 2026
Advisor,

(Arrie Kurniawardhani, S.Si., M.Kom.)

EXAMINER ENDORSEMENT PAGE

Multi-Sensor Fusion for Tropical Forest Carbon Stock Prediction

THESIS

Has been defended in front of the examiners as one of the requirements to obtain a Bachelor of Informatics degree from the Undergraduate Program in Informatics at the Faculty of Industrial Technology, Universitas Islam Indonesia
Yogyakarta, January 9th, 2026

Chair

Arrie Kurniawardhani, S.Si., M.Kom. _____

Examiner 1

Ir. Irving Vitra Paputungan, S.T., M.Sc., Ph.D. _____

Examiner 2

Dr. Nur Wijayaning Rahayu, S.Kom., M.Cs. _____

Acknowledged by,
Head of Undergraduate Program in Informatics
Faculty of Industrial Technology
Universitas Islam Indonesia

(Ir. Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D.)

AUTHENTICITY STATEMENT

The undersigned:

Name	:	Vladimirrahman Salsabil Abdullah
Student ID	:	22523298

Final project with title:

Multi-Sensor Fusion for Tropical Forest Carbon Stock Prediction

Stating that all components and contents in this final project are my own work. If in the future it is proven that some parts of this work are not my own work, the final project submitted as my own work is ready to be withdrawn and ready to bear any risks and consequences.
Thus this statement letter is made, hopefully it can be used properly.

Yogyakarta, January 07, 2026

(Vladimirrahman Salsabil Abdullah)

DEDICATION

This thesis is dedicated to:

1. My late father, Emillano Suryarama Triputra To his loving memory. His values, strength, and guidance continue to shape who I am today. Although he is no longer here to witness this achievement, his spirit, sacrifices, and unwavering belief in the importance of education remain my greatest source of inspiration. This work stands as a tribute to his legacy and to the lessons he instilled, which continue to guide my journey.
2. My mother, Santi Rahmawati For her unwavering love, strength, and sacrifices, which have formed the foundation of my academic journey. Her constant encouragement, patience, and belief in my abilities provided me with the resilience to overcome challenges and remain committed to this research. Without her guidance and support, this achievement would not have been possible.
3. My family For their continuous support, patience, and encouragement, which played an essential role in the completion of this work.
4. My Thesis Supervisor, Arrie Kurniawardhani, S.Si., M.Kom. For her invaluable supervision, insightful guidance, and consistent support throughout the research process.
5. My Academic Supervisor, Dr. Nur Wijayaning Rahayu, S.Kom., M.Cs. For her guidance and constructive input during the early stages of my academic development.
6. My partner, Mita Andriani For her unwavering support, understanding, and encouragement, which provided emotional strength and motivation throughout the research journey.
7. My close friends, Abdullah Hisyam, Muhammad Rifqi, Muhammad Rakha Savero Zulni, Khalil Khalabi, Raja Fawwaz Aushaf, Wisnu Arya Pradipta, A'rafi Laksmana Dirgantara, Muhammad Fazli Ramadhani Sukma, Muhammad Aulia Nalendra, Muhammad Ferrel Ganendra Arisaputra, Khun Muhammad Dalle Jasmin and Muhammad Fiqri For their companionship, thoughtful discussions, and shared experiences, which offered balance and resilience during demanding periods of this work.

8. My favorite football team Manchester United For their enduring spirit, resilience, and pursuit of excellence, which served as a personal source of inspiration during challenging moments.

MOTTO

Hope for the Best, Prepare for the Worst

FOREWORD

Praise be to Allah SWT, the Most Gracious and the Most Merciful, for His blessings and guidance, which have enabled me to complete this thesis titled “Multi-Sensor Fusion for Tropical Forest Carbon Stock Prediction.”

This thesis is submitted as a partial fulfillment of the requirements for the degree of Bachelor of Informatics at Universitas Islam Indonesia.

The completion of this research would not have been possible without the assistance, support, and guidance of many individuals. On this occasion, I would like to express my deepest gratitude to:

1. Prof. Fathul Wahid, S.T., M.Sc., Ph.D., the Rector of Universitas Islam Indonesia, for the opportunity to pursue my studies at this esteemed university.
2. Prof. Dr. Ir. Hari Purnomo., M.T., IPU, ASEAN.Eng, the Dean of the Faculty of Industrial Technology, for providing the academic facilities that supported my learning journey.
3. Raden Teduh Dirgahayu, S.T., M.Sc., the Head of the Department of Informatics, for the support and administration of the department.
4. Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D., the Head of the Study Program of Undergraduate Informatics, for the guidance and management of the academic curriculum.
5. Arrie Kurniawardhani, S.Si., M.Kom., my Thesis Supervisor, for her patience, invaluable supervision, and insightful guidance in navigating the complexities of multi-sensor fusion and machine learning models throughout this research.
6. Dr. Nur Wijayaning Rahayu, S.Kom., M.Cs., my Academic Supervisor, for her advice and direction during my study period.
7. All lecturers and staff of the Department of Informatics for the knowledge shared and assistance provided during my studies.

I also extend my heartfelt thanks to my parents, family, and friends, whose names are dedicated on the previous page, for their endless love and moral support.

I realize that this thesis is far from perfect. Therefore, I welcome any constructive criticism and suggestions for future improvements. I hope this research on tropical forest carbon stock estimation provides benefits for the development of science and contributes to climate change mitigation efforts.

Yogyakarta, January 07, 2026

(Vladimirrahman salasbil Abdullah)

ABSTRACT

Tropical forests store approximately 25% of terrestrial carbon, making accurate quantification of Aboveground Biomass Density (AGBD) critical for global climate mitigation and national reporting under REDD+ mechanisms. However, operational biomass mapping in dense tropical environments faces significant challenges due to signal saturation in optical and C-band Synthetic Aperture Radar (SAR) sensors. While recent studies have explored neural networks for biomass estimation, many rely on incomplete sensor stacks that fail to capture the complex structure of mature forests.

This study proposes a robust Multi-Sensor Data Fusion framework to predict AGBD in the Special Region of Yogyakarta, integrating Sentinel-1 (C-band SAR), Sentinel-2 (Optical), and ALOS PALSAR-2 (L-band SAR) imagery. Calibrated estimates from the Global Ecosystem Dynamics Investigation (GEDI) Level 4A product were utilized as high-fidelity reference data. To identify the optimal modeling strategy, sixteen experimental configurations were rigorously evaluated, testing four feature selection algorithms (Recursive Feature Elimination, Mutual Information, SelectKBest, and PCA) paired with four regression models (Linear Regression, Random Forest, SVR, and Multi-Layer Perceptron).

The experimental results demonstrate that the proposed fusion architecture, utilizing Recursive Feature Elimination (RFE) with a Multi-Layer Perceptron (MLP), achieved the most robust performance, yielding an R^2 of 0.3432, an RMSE of 74.02 Mg/ha, and a Mean Absolute Percentage Error (MAPE) of 69.98%. While the Support Vector Regression (SVR) model yielded a lower percentage error (MAPE 54.41%), it was rejected due to a systematic negative bias (-14.25 Mg/ha), which resulted in the severe underestimation of high-carbon stock areas. Conversely, univariate filter methods (Mutual Information) proved ineffective ($R^2 < 0.26$) as they prioritized optical indices that saturate early, discarding essential L-band volume scattering information.

This thesis confirms that incorporating L-band SAR is scientifically necessary to overcome the saturation limits of C-band-only approaches. The resulting model provides a scalable, scientifically justified method for high-resolution carbon stock mapping, supporting Indonesia's commitments to the Paris Agreement.

Keywords: Carbon Stock, Multi-Sensor Fusion, Random Forest, Multi-Layer Perceptron, GEDI LiDAR, Tropical Forest.

GLOSSARY

Term	Definition
Above-Ground Biomass (AGB)	The total amount of living organic matter above the soil surface, including stems, branches, bark, seeds, and foliage; expressed in Mg/ha.
Backpropagation	A learning algorithm in neural networks that calculates loss gradients and propagates errors backward to update model weights.
Carbon Stock	The total quantity of carbon stored in vegetation or soil; in this study derived from AGB using a carbon fraction of 0.47.
Epoch	One complete pass of the entire dataset through a neural network during training.
Feature Extraction	The process of transforming raw data into useful variables, such as spectral indices or radar texture metrics.
GEDI (Global Ecosystem Dynamics Investigation)	A spaceborne LiDAR instrument that provides high-resolution structural and biomass measurements of forests.
GLCM (Gray Level Co-occurrence Matrix)	A statistical method for extracting texture features (e.g., Contrast, Entropy, Correlation) from imagery.
Hyperparameter	A model configuration variable (e.g., number of trees, learning rate) that cannot be learned from data.
L-Band	A radar frequency band (~1.27 GHz) used by ALOS PALSAR, capable of penetrating forest canopies.
LiDAR (Light Detection and Ranging)	A remote sensing technology using laser pulses to measure distances and generate 3D information.
Multi-Layer Perceptron (MLP)	A feedforward neural network with one or more hidden layers capable of modeling non-linear relationships.
Multi-Sensor Fusion	The integration of data from multiple sensors (e.g., Optical, Radar, LiDAR) to improve accuracy.
NDVI (Normalized Difference Vegetation Index)	A vegetation index derived from the NIR and Red bands to assess plant health and density.
Random Forest (RF)	An ensemble learning method that builds multiple decision trees and averages their predictions.
Recursive Feature Elimination (RFE)	A feature selection method that removes the weakest features iteratively to find the optimal subset.
REDD+	A global framework to reduce emissions from deforestation and forest degradation and enhance forest carbon stocks.
ReLU (Rectified Linear Unit)	A neural network activation function defined as $f(x) = max(0, x)$.
RMSE (Root Mean Square Error)	A regression performance metric representing the standard deviation of prediction errors.
Saturation	A remote sensing issue where increases in biomass no longer produce proportional increases in sensor response.
Sentinel-1	A C-band SAR mission by ESA providing all-weather radar imagery.

Sentinel-2	An ESA optical mission providing high-resolution multispectral imagery.
SRTM (Shuttle Radar Topography Mission)	A mission providing near-global elevation data for high-resolution digital elevation models.
Synthetic Aperture Radar (SAR)	An active radar system that transmits microwave pulses and measures backscatter to produce high-resolution images.

TABLE OF CONTENTS

TITLE PAGE i
SUPERVISOR ENDORSEMENT PAGE ii
EXAMINER ENDORSEMENT PAGE iii
AUTHENTICITY STATEMENT iv
DEDICATION v
MOTTO vii
FOREWORD viii
ABSTRACT x
GLOSSARY xi
TABLE OF CONTENTS xiii
LIST OF TABLE xvi
LIST OF FIGURES xvii
CHAPTER I INTRODUCTION 1
1.1 Background of Study 1
1.2 Problem Statement 2
1.3 Research Question 2
1.4 Research Objectives 2
1.5 Significance of the Study 3
1.6 Scope and Limitations 3
1.6.1 Scope of the Study 3
1.6.2 Limitations of the Study 4
1.7 General Research Methodology 4
1.7.1 Literature Review and Theoretical Exploration 5
1.7.2 Problem Identification and Objective Formulation 5
1.7.3 Data Acquisition and Pre-processing 5
1.7.4 Methodological Development and Implementation 6
1.7.5 Result Analysis and Validation 6
1.8 Report Structure 6
CHAPTER II LITERATURE REVIEW 8
2.1 Theoretical Framework of Forest Biomass 8
2.1.1 Definition of AGB and Carbon Stock 8
2.1.2 The Importance of Carbon Stock Monitoring 9
2.1.3 Biomass and Carbon Stock Classification 9
2.2 Remote Sensing for Biomass Estimation 10
2.2.1 Optical Remote Sensing 10
2.2.2 Radar Remote Sensing 12
2.2.3 LiDAR Remote Sensing (GEDI) 12
2.2.4 Sentinel – 1 13
2.2.5 Sentinel – 2 14
2.2.6 ALOS PALSAR – 2 16
2.2.7 GEDI 17
2.3 Multi-Sensor Data Fusion 18
2.3.1 The Rationale for Synergistic Fusion 18
2.3.2 Review of GEDI Fusion Studies 19
2.4 Feature Selection 20
2.4.1 Recursive Feature Elimination 20
2.4.2 Mutual Information 21
2.4.3 SelectKBest with ANOVA F-test 22

2.4.4 Principal Component Analysis....23
2.5 Machine Learning in Remote Sensing....24
2.5.1 Random Forest....24
2.5.2 Multi-Layer Percepton....25
2.5.3 Support Vector Regression....27
2.5.4 Multi Linear Regression....29
CHAPTER III RESEARCH METHODS....32
3.1 Study Area....33
3.2 Data Acquisition....34
3.2.1 GEDI Level 4A Biomass Data (Reference Data)....36
3.2.2 Sentinel-2 MSI Imagery (Optical)....37
3.2.3 Sentinel-1 SAR Imagery (C-Band Radar)....38
3.2.4 ALOS PALSAR-2 (L-Band Radar)....38
3.2.5 Topographic Data (SRTM)....38
3.2.6 Land Cover Masking (ESA WorldCover)....39
3.3 Data Pre-processing and Feature Extraction....39
3.3.1 Optical Processing and Spectral Indices....39
3.3.2 Radar Processing and Texture Analysis....40
3.4 Master Dataset Creation....41
3.5 Feature Selection....41
3.5.1 Feature Selection (RFE)....41
3.5.2 Mutual Information....42
3.5.3 Select KBest (F-Test)....42
3.5.4 Principal Component Analysis....43
3.6 Machine Learning Models....43
3.6.1 Random Forest....43
3.6.2 Multi-Layer Perceptron....44
3.6.3 Support Vector Regression....47
3.6.4 Multiple Linear Regression....48
3.7 Model Evaluation....48
3.8 Alternative Workflow in the Absence of L-Band SAR Data....50
3.8.1 Rationale for an L-Band–Independent Workflow....50
3.8.2 Modified Data Sources and Predictor Set....51
3.8.3 Feature Engineering Adjustment....52
3.8.4 Feature Selection Strategy Without L-Band SAR....52
3.8.5 Model Training and Validation Configuration....54
3.9 Final Inference....55
3.9.1 Applying the Best Model....55
3.9.2 Generating Spatial Carbon Stock Map....55
CHAPTER IV RESULT....56
4.1 Feature Selection Result....56
4.1.1 Mutual Information....56
4.1.2 Recursive Feature Elimination....56
4.1.3 SelectKBest (Anova F-Test)....56
4.1.4 Principal Component Analysis....56
4.1.5 Consensus Feature Set....57
4.1.6 Comparison of Feature Sets....57
4.2 Model Performance Across All Feature Selection and Machine Learning Combination....58
4.2.1 Cross-Validation & Training Results....59

4.2.2 Best Performing Model 60
4.3 Model Testing 61
4.3.1 Evaluation of Predicted vs. Observed Biomass 61
4.3.2 Summary of Testing Analysis 65
4.4 Model Performance Withoud L-Band SAR Data 66
4.4.1 Objective of the L-Band Exclusion Experiment 66
4.4.2 Feature Selection Without L-Band SAR 66
4.4.3 Model Performance Comparison 68
4.4.4 Best Performing Model Without L-Band SAR 69
4.5 Final Inference Using the Best Model 71
4.5.1 Generation of the Spatial Carbon Stock Map 71
4.5.2 Spatial Distribution Analysis 71
4.6 Discussion 73
4.6.1 Efficacy of Feature Selection Strategies 73
4.6.2 Neural Networks vs. Traditional Regression 73
4.6.3 The Role of Multi-Sensor Fusion 74
4.6.4 Uncertainties and the Impact of GEDI Noise 74
CHAPTER V CONCLUSION 75
5.1 Conclusion 75
5.2 Limitations of the Study 75
5.3 Recommendations 76
REFERENCE 77
APPENDIX 81

LIST OF TABLE

Table 3.1 Summary of satellite sensor specifications, temporal coverage, and spatial resolution used in the study.....35
Table 3.2 Mathematical formulations of the spectral indices used for biomass feature extraction.....40
Table 4.1 Performance metrics for all Feature Selection and Machine Learning combinations. 59
Table 4.2 Performance metrics for Sentinel-only model combinations (No L-Band). 68
Table 4.3 Performance degradation analysis due to L-band exclusion 70

LIST OF FIGURES

Figure 2.1 Architecture of a feedforward neural network with an input layer, one hidden layer, and an output layer. 26
Figure 3.1 Workflow for Predicting mapping 32
Figure 3.2 Map of the study area showing the administrative boundaries of the Special Region of Yogyakarta and the distribution of forest cover types. 33
Figure 3.3 Spatial distribution of valid GEDI L4A footprints across the Special Region of Yogyakarta after quality filtering 37
Figure 3.4 Code Snippet: Standardization of Training and Test Features Using Z-Score Normalization. 45
Figure 3.5 Code Snippet : Model implementation 46
Figure 4.1 Scatter Plot for MLP with RFE 62
Figure 4.2 Scatter Plot for MLP with PCA 63
Figure 4.3 Scatter Plot for MLP with MI 64
Figure 4.4 Scatter Plot for RF with RFE 65
Figure 4.5 illustrate the predicted carbon stock distribution for the year 2020 using the MLP. 72

CHAPTER I
INTRODUCTION

1.1 Background of Study

Tropical forests serve as a cornerstone of the global terrestrial carbon cycle, acting as massive carbon sinks that mitigate anthropogenic climate change. It is estimated that these ecosystems store approximately 25% of terrestrial carbon and account for over 30% of global net primary production (Prada et al., 2025). Consequently, the accurate quantification of Above-Ground Biomass (AGB) is a strict prerequisite for international climate policies. Under frameworks such as REDD+ and the Paris Agreement, nations are mandated to report carbon stocks with high precision to verify Nationally Determined Contributions (NDCs) (Butler et al., 2024).

Historically, biomass estimation relied on field inventories. While accurate at the plot level, this method is labor-intensive, spatially discontinuous, and often logistically impossible in remote terrain. To achieve the landscape-scale monitoring required for national reporting, satellite remote sensing has become the operational standard. However, relying on single-sensor approaches presents significant technical limitations. Optical sensors like Sentinel-2 are hampered by persistent cloud cover and signal saturation in high-biomass areas (Abbas et al., 2020). Similarly, C-band Synthetic Aperture Radar (SAR), such as Sentinel-1, often saturates at low biomass levels due to limited canopy penetration.

The recent availability of spaceborne LiDAR, specifically the Global Ecosystem Dynamics Investigation (GEDI), has revolutionized this field by providing direct measurements of vertical forest structure—the strongest predictor of biomass (Ni-Meister et al., 2025). Yet, GEDI provides only discrete samples along ground tracks, failing to produce the continuous "wall-to-wall" maps required by policymakers. Therefore, scientific consensus has shifted toward multi-sensor data fusion. By integrating the structural precision of GEDI with the spatial continuity of optical and radar imagery, researchers can model relationships that overcome individual sensor limitations (Yu et al., 2025).

To overcome these limitations, this research employs a multi-sensor fusion framework that integrates Sentinel-1 SAR, Sentinel-2 optical imagery, and spaceborne LiDAR from GEDI to generate spatially continuous and structurally informed predictors of forest carbon. By applying advanced regression modeling techniques, this study aims to develop an accurate and scalable approach for predicting Above-Ground Biomass (AGB) in tropical forests. Rather than

emphasizing model benchmarking, the primary objective is to construct a reliable predictive model capable of producing high-resolution carbon stock estimates that support national reporting requirements and climate mitigation policies.

1.2 Problem Statement

Accurately predicting tropical forest carbon stocks remains challenging despite the rapid expansion of satellite-based monitoring technologies. Optical and C-band radar sensors such as Sentinel-2 and Sentinel-1 continue to suffer from signal saturation, limiting their ability to characterize high-biomass forests—the areas most critical for REDD+ reporting. Meanwhile, GEDI LiDAR provides highly accurate structural information but only at discrete footprints, preventing direct wall-to-wall biomass mapping. As a result, current approaches often produce incomplete or biased estimates, particularly in mature, carbon-dense forests. Therefore, there is an urgent need for a multi-sensor fusion framework that integrates the complementary strengths of optical, radar, and LiDAR data to generate reliable, high-resolution, and spatially continuous predictions of Above-Ground Biomass across heterogeneous tropical landscapes.

1.3 Research Question

Based on the identified technical gaps and the necessity for improved regional carbon monitoring, this study addresses the following key questions:

1. How can multi-sensor data—integrating optical (Sentinel-2), C-band SAR (Sentinel-1), L-band SAR (e.g., ALOS PALSAR-2), and GEDI LiDAR—be fused to generate reliable and spatially continuous predictions of Above-Ground Biomass in tropical forests?
2. Which variables derived from optical, C-band SAR, L-band SAR, and LiDAR contribute most significantly to improving the accuracy of carbon stock prediction across heterogeneous tropical landscapes?
3. To what extent can a regression-based modeling framework capture the complex, non-linear relationships among multi-sensor inputs (optical, C-band, L-band, LiDAR) for producing high-resolution carbon stock maps?

1.4 Research Objectives

To address the research questions outlined above, this study aims to achieve the following specific objectives:

1. To develop a multi-sensor fusion framework that integrates optical (Sentinel-2), C-band SAR (Sentinel-1), L-band SAR (ALOS PALSAR-2), and GEDI LiDAR data for generating spatially continuous predictions of Above-Ground Biomass in tropical forests.
2. To identify and evaluate the relative importance of predictor variables derived from optical imagery, C-band SAR, L-band SAR, and LiDAR metrics in improving the accuracy of carbon stock prediction.
3. To build a regression-based modeling approach capable of capturing the non-linear relationships between multi-sensor inputs and field-estimated biomass to produce high-resolution carbon stock maps.

1.5 Significance of the Study

This study provides an important contribution to tropical forest carbon monitoring by advancing a multi-sensor fusion approach that integrates optical, C-band SAR, L-band SAR, and GEDI LiDAR data for accurate and spatially continuous prediction of Above-Ground Biomass (AGB). By addressing the long-standing limitations of single-sensor methods—such as signal saturation in high-biomass areas and the discontinuous sampling of spaceborne LiDAR—this research supports the development of carbon mapping techniques that are both operationally feasible and scientifically robust. The resulting carbon stock predictions have direct relevance for national reporting under REDD+ and the Paris Agreement, enabling countries to comply with increasingly stringent transparency and accuracy requirements. Furthermore, the study enhances understanding of the contributions of different sensor modalities to biomass estimation, offering methodological insights for future remote sensing applications in forest monitoring. Beyond academic value, this research provides practical benefits for policymakers, conservation agencies, and land managers seeking reliable, high-resolution carbon data to guide climate mitigation strategies, forest protection efforts, and sustainable land-use planning in tropical regions.

1.6 Scope and Limitations

1.6.1 Scope of the Study

This study focuses on predicting tropical forest Above-Ground Biomass (AGB) through the integration of multi-sensor remote sensing data. The analysis is limited to the fusion of optical imagery (Sentinel-2), C-band SAR (Sentinel-1), L-band SAR (ALOS PALSAR-2), and spaceborne LiDAR measurements from GEDI. The study area is restricted to tropical forest

landscapes within the selected region in Indonesia, and all modeling efforts pertain specifically to this geographic context. Field plot data used for training and validation rely on existing ground-based measurements available for the study area. The research is confined to regression-based modeling for biomass prediction and does not include algorithm benchmarking or performance comparison as a primary objective. Additionally, the study focuses solely on Above-Ground Biomass and does not address below-ground or soil carbon pools. Temporal analysis, forest degradation detection, and change monitoring are outside the scope of this work. The final outputs are limited to static, high-resolution AGB maps generated for the selected study period.

1.6.2 Limitations of the Study

Several limitations influence the outcomes of this study. First, the accuracy of the biomass prediction model is constrained by the availability and quality of field plot data, which may not fully represent the variability of tropical forest conditions across the study area. Second, GEDI LiDAR data provide precise structural information but are limited by their sampling footprint distribution, which may leave large gaps in areas with sparse coverage. Third, the optical and radar datasets used in this research are affected by inherent sensor limitations, including cloud contamination in Sentinel-2 imagery, moisture sensitivity in Sentinel-1 backscatter, and temporal inconsistencies between acquisition dates of multi-sensor inputs. Fourth, the L-band SAR data (ALOS PALSAR-2) may differ in acquisition period relative to GEDI and optical images, potentially introducing temporal mismatches that influence model accuracy. Fifth, the modeling framework focuses solely on regression-based approaches and does not evaluate all existing or emerging machine learning methods, which may limit the achievable predictive performance. Lastly, the study is geographically restricted to a selected tropical forest region in Indonesia, which may reduce the generalizability of the findings to other forest types or biogeographical contexts.

1.7 General Research Methodology

This research follows a systematic scientific workflow to ensure the validity, reproducibility, and academic rigor of the study. Unlike the technical implementation details discussed in Chapter III, which focus on the specific algorithms and code architecture, this section outlines the operational phases of the research lifecycle, ranging from the initial problem identification and literature review to the final analysis and report generation.

The research process is divided into six distinct phases, elaborated below:

1.7.1 Literature Review and Theoretical Exploration

The research commenced with a comprehensive literature review conducted between October 2025 and November 2025. The objective was to understand the state-of-the-art methods in forest biomass estimation and identify the specific limitations of single-sensor approaches.

- Data Sources: The review utilized academic databases including IEEE Xplore, ScienceDirect, MDPI, and Google Scholar.
- Search Keywords: The primary search strings employed were: "Multi-sensor data fusion for biomass," "GEDI LiDAR applications," "Sentinel-1 and Sentinel-2 forest carbon," "Deep Learning for remote sensing," and "ALOS PALSAR-2 saturation."
- Selection Criteria: References were prioritized based on recency (published within the last 5–7 years) and relevance to tropical forest ecosystems. This phase established the theoretical framework for using GEDI L4A as a reference dataset and justified the selection of the Multi-Layer Perceptron (MLP) as the primary modeling architecture.

1.7.2 Problem Identification and Objective Formulation

Based on the literature review, the specific gap addressed in this study was identified: the "saturation phenomenon" in optical and C-band radar sensors that leads to the underestimation of high-biomass tropical forests. Consequently, the research questions and objectives (as outlined in Sections 1.3 and 1.4) were formulated to focus on determining whether the inclusion of L-band SAR and non-linear deep learning models could overcome this limitation in the Special Region of Yogyakarta.

1.7.3 Data Acquisition and Pre-processing

This phase involved the collection of remote sensing data required for the study. Data acquisition was performed entirely via the Google Earth Engine (GEE) platform to ensure computational efficiency.

- Target Data: GEDI Level 4A biomass estimates were filtered for the year 2020.
- Predictor Data: Satellite imagery from Sentinel-2 (Optical), Sentinel-1 (C-band Radar), ALOS PALSAR-2 (L-band Radar), and SRTM (Topography) was acquired.

- Constraints: All data were restricted to the 2020 dry season (June–September) to minimize cloud cover and soil moisture interference.

1.7.4 Methodological Development and Implementation

This phase represents the core technical execution of the research. It encompassed the creation of the "Master Dataset" by extracting pixel values at GEDI footprint locations. Various computational experiments were designed, specifically:

1. Feature Selection: Implementing algorithms like Recursive Feature Elimination (RFE) to reduce data dimensionality.
2. Model Training: Developing and tuning four distinct regression models (MLR, RF, SVR, and MLP) using Python libraries (Scikit-Learn).
3. Ablation Study: Designing a specific experiment to test model performance with and without L-band SAR data to simulate data-scarce scenarios.

1.7.5 Result Analysis and Validation

Following model training, the results were rigorously evaluated using statistical metrics. The analysis focused on comparing predicted values against observed GEDI biomass using an independent test set (20% of data). Key activities included:

- Calculating metrics such as R^2, RMSE, and Bias.
- Generating scatter plots to visualize the "saturation" effect.
- Producing the final wall-to-wall Carbon Stock Map of the Special Region of Yogyakarta for spatial analysis.

1.8 Report Structure

This thesis is organized into five distinct chapters, systematically guiding the reader from the theoretical foundation to the empirical findings and final conclusions. The structure of the report is outlined as follows:

- Chapter I: Introduction This chapter establishes the context of the study. It outlines the background of the research, highlighting the importance of tropical forests in the global carbon cycle and the limitations of current monitoring methods. It defines the problem statement regarding signal saturation in single-sensor approaches and formulates the specific research questions and objectives. The chapter concludes with the significance of the study, the scope of the research, and the limitations faced during execution.

- Chapter II: Literature Review This chapter provides the theoretical framework necessary to understand the research. It reviews the fundamental concepts of Above-Ground Biomass (AGB) and Carbon Stock. It critically analyzes the physical principles of the remote sensing technologies employed, including Optical (Sentinel-2), C-band Radar (Sentinel-1), L-band Radar (ALOS PALSAR-2), and LiDAR (GEDI). Furthermore, it explores existing studies on multi-sensor fusion and discusses the theoretical basis for the Feature Selection algorithms (RFE, Mutual Information, PCA) and Machine Learning models (Random Forest, MLP, SVR) utilized in this study.
- Chapter III: Research Methods This chapter details the technical workflow implemented to achieve the research objectives. It describes the study area in the Special Region of Yogyakarta and the data acquisition specifications for the 2020 dry season. The methodology elaborates on data pre-processing, feature extraction (spectral indices and GLCM textures), and the creation of the Master Dataset. It further explains the configuration of the machine learning algorithms and the statistical metrics used for model evaluation. Finally, it outlines the experimental design for testing model performance in the absence of L-band SAR data.
- Chapter IV: Results and Discussion This chapter presents the empirical findings of the study. It begins by analyzing the outcomes of the different feature selection strategies. It then provides a comparative performance analysis of the sixteen model configurations, identifying the optimal architecture. The chapter includes a rigorous validation using scatter plots to visualize prediction accuracy and saturation effects. Additionally, it presents the results of the L-band exclusion experiment and displays the final spatial distribution map of carbon stocks in Yogyakarta. The discussion interprets these findings in the context of sensor fusion theory and ecological significance.
- Chapter V: Conclusion The final chapter summarizes the key findings, confirming the superiority of the Multi-Layer Perceptron (MLP) model and the critical role of L-band radar. It reiterates the limitations of the study, such as reliance on GEDI reference data , and offers recommendations for future research and operational implementation to support national carbon reporting.

CHAPTER II
LITERATURE REVIEW

2.1 Theoretical Framework of Forest Biomass

The estimation of forest biomass serves as the fundamental basis for understanding the global carbon cycle. This section establishes the theoretical definitions utilized in this study and delineates the critical importance of monitoring forest carbon stocks in the context of climate change mitigation.

2.1.1 Definition of AGB and Carbon Stock

Forest biomass is generally categorized into two primary components: Above-Ground Biomass (AGB) and Below-Ground Biomass (BGB). For the purposes of remote sensing applications, Above-Ground Biomass (AGB) is defined as the total amount of living organic matter found above the soil surface, inclusive of the stem, stump, branches, bark, seeds, and foliage (Food and Agriculture Organization of the United Nations, 2010). AGB is typically expressed in units of dry weight per unit area, such as megagrams per hectare (Mg/ha) or tons per hectare (t/ha).

Carbon Stock refers to the absolute quantity of carbon held within a specific pool (e.g., vegetation or soil) at a specified time. In forest ecology, there is a direct biophysical relationship between biomass and carbon content. Plants absorb atmospheric carbon dioxide (CO_2) through photosynthesis and store it as biomass. The Intergovernmental Panel on Climate Change (IPCC) guidelines suggest that approximately 47% to 50% of dry wood biomass consists of carbon. Therefore, a widely accepted conversion factor of 0.47 (or 0.5 for simplification) is applied to convert AGB estimates into Carbon Stock (Doraisami et al., 2024).

$$Carbon\ Stock = AGB \times CF \tag{2.1}$$

In equation 2.1, Carbon Stock (expressed in Mg C/ha) represents the final estimated mass of elemental carbon stored within the forest vegetation. AGB (Mg/ha) denotes the Above-Ground Biomass, quantifying the total dry mass of all living organic matter located above the soil surface. The variable CF refers to the Carbon Fraction, a biophysical constant representing the proportion of elemental carbon contained within dry biomass. For the context of this study, a coefficient of 0.47 is utilized, following the standard recommendation for tropical forest ecosystems (Doraisami et al., 2024).

2.1.2 The Importance of Carbon Stock Monitoring

Monitoring forest carbon stocks is not merely a technical exercise but a critical requirement for global climate policy and environmental management.

1. Climate Change Mitigation: Tropical forests act as a significant global carbon sink, sequestering gigatons of anthropogenic carbon emissions. Accurate monitoring is essential to quantify this sequestration capacity. Uncertainty in biomass estimates leads to uncertainty in global carbon budget calculations, affecting predictions of future climate scenarios (Mo et al., 2023)
4. REDD+ Implementation: The mechanism for Reducing Emissions from Deforestation and Forest Degradation (REDD+) relies heavily on Measurement, Reporting, and Verification (MRV) systems. For developing nations like Indonesia to receive result-based payments for conservation, they must provide transparent, consistent, and accurate estimates of their forest carbon baselines (Praputra et al., 2016; Yang et al., 2025).
5. Biodiversity and Ecosystem Health: High biomass levels are often correlated with structural complexity and species diversity (Mansingh et al., 2025). Thus, mapping carbon stocks often serves as a proxy for identifying high-value conservation areas and monitoring forest degradation that might not be visible through simple forest-cover change detection (Soto-Navarro et al., 2020).

2.1.3 Biomass and Carbon Stock Classification

To interpret the ecological significance of the predicted values, this study adopts a biomass classification framework informed by the guidelines and default assumptions of the Intergovernmental Panel on Climate Change (IPCC).

Biologically, carbon stock is commonly estimated as a fraction of Aboveground Biomass (AGB), using a conversion factor of 0.47 (IPCC, 2006). Based on tropical rainforest structural characteristics and ranges commonly reported in the literature, aboveground biomass is categorized into three tiers:

1. Low Carbon Stock (< 100 Mg/ha AGB): Represents degraded lands, shrublands, or young regenerating forests. These areas are characterized by simplified stand structure, limited canopy development, and reduced carbon sequestration capacity.

2. Moderate Carbon Stock (100–200 Mg/ha AGB): Corresponds to secondary forests or agroforestry systems. These ecosystems are typically in a recovery phase, exhibiting relatively high carbon accumulation rates but lower biomass compared to old-growth forests.
3. High Carbon Stock (> 200 Mg/ha AGB): Represents primary forests or mature secondary forests. Tropical forests exceeding this threshold function as major carbon sinks and are widely recognized as high-conservation-value ecosystems critical for climate change mitigation.

Therefore, within the context of this study, predicted AGB values exceeding 140–150 Mg/ha are considered ecologically significant, indicating a structural transition from secondary regrowth toward mature forest conditions.

2.2 Remote Sensing for Biomass Estimation

Remote sensing technology has revolutionized the monitoring of forest ecosystems, offering a scalable alternative to labor-intensive field inventories. This section reviews the physical principles, advantages, and limitations of the three primary modalities utilized in this study: Optical, Synthetic Aperture Radar (SAR), and Light Detection and Ranging (LiDAR).

2.2.1 Optical Remote Sensing

Optical remote sensing relies on passive sensors that measure reflected solar radiation from the Earth's surface. In the context of vegetation monitoring, the fundamental principle is based on the spectral signature of healthy plants: strong absorption of red light by chlorophyll pigments for photosynthesis and high reflectance of near-infrared (NIR) energy due to the internal cellular structure of leaves.

To quantify this spectral behavior, Vegetation Indices (VIs) are mathematically derived to maximize sensitivity to vegetation parameters while minimizing atmospheric and soil background noise.

Vegetation Indices (NDVI and EVI) The Normalized Difference Vegetation Index (NDVI) is the most widely utilized metric for assessing vegetation health and density. It is calculated as the normalized difference between the NIR and Red bands:

$$NDVI = \frac{\rho NIR - \rho RED}{\rho NIR + \rho RED} \tag{2.2}$$

Equation 2.2 quantifies the Normalized Difference Vegetation Index (NDVI) by calculating the ratio between the Near-Infrared (ρNIR) and Red (ρRED) spectral bands. This mathematical relationship exploits the distinct biological properties of healthy vegetation, which strongly absorbs red light for photosynthesis while reflecting near-infrared energy due to leaf cell structure. The resulting value provides a standardized measure of chlorophyll content and vegetation vigor, though it becomes less sensitive as biomass increases.

While robust for general cover mapping, NDVI is sensitive to soil brightness and atmospheric effects. To address these limitations, the Enhanced Vegetation Index (EVI) was developed, as expressed in Equation 2.3:

$$EVI = G \times \frac{\rho NIR - \rho RED}{\rho NIR + C_1 \times \rho RED - C_2 \times \rho BLUE + L} \tag{2.3}$$

Equation 2.3 details the Enhanced Vegetation Index (EVI), an optimized formula designed to decouple the vegetation signal from background noise. In this equation, $\rho BLUE$ represents the blue band reflectance which, combined with coefficients $C_1$and C_2, corrects for atmospheric aerosol scattering. Additionally, the variable L serves as a canopy background adjustment factor, and G acts as a gain factor. This structural complexity allows EVI to maintain sensitivity in high-biomass regions where NDVI typically saturates, making it particularly valuable for tropical forest analysis.

The Saturation Problem A critical limitation of optical remote sensing for biomass estimation is the "saturation phenomenon." In complex, multi-layered tropical forests, optical sensors primarily capture the spectral response of the upper canopy closure. Once the canopy closes completely (typically around 100–150 Mg/ha), the relationship between spectral reflectance and biomass becomes asymptotic. The sensor cannot "see" the accumulation of woody biomass in the trunks and branches beneath the leaves, leading to significant underestimation of AGB in dense forests (Chen et al., 2025).

2.2.2 Radar Remote Sensing

Unlike optical sensors, Synthetic Aperture Radar (SAR) is an active system that transmits microwave pulses and records the backscattered energy. This capability allows SAR to operate day and night and penetrate cloud cover—a crucial advantage in tropical regions like Indonesia.

Principles of Backscatter and Wavelength The intensity of the radar backscatter (σ^0) is influenced by the sensor's wavelength (λ^0), polarization, and the target's dielectric constant and surface roughness. The interaction between the radar signal and forest structure is governed by the wavelength:

- Short Wavelengths (X and C bands): Primarily interact with small structural elements like leaves and twigs in the upper canopy.
- Long Wavelengths (L and P bands): Have the physical capacity to penetrate through the canopy gaps and interact with larger structural components like main branches and trunks, where the majority of biomass is stored.

Review of Sentinel-1 (C-Band) The European Space Agency's (ESA) Sentinel-1 constellation operates in the C-band frequency (~5.405 GHz, $\lambda \approx 5.6$ cm). While Sentinel-1 offers high temporal resolution and free accessibility, its short wavelength limits its utility for high-biomass estimation. C-band signals are typically scattered by the upper canopy foliage, resulting in rapid signal saturation at relatively low biomass levels (often <100 Mg/ha). However, Sentinel-1 remains valuable for texture analysis and identifying recent forest disturbances.

Review of ALOS PALSAR (L-Band**)** The ALOS PALSAR systems (PALSAR-1 and PALSAR-2), developed by JAXA, operate in the L-band frequency (~1.27 GHz, $\lambda \approx 23.6$ cm). The longer L-band wavelength penetrates the forest canopy, interacting directly with the stems and trunks via double-bounce scattering mechanisms. This physical interaction results in a strong, linear correlation with AGB up to much higher density levels compared to optical or C-band sensors (Yu & Saatchi, 2016). Consequently, L-band SAR is widely regarded as the gold standard for operational spaceborne biomass mapping in the tropics.

2.2.3 LiDAR Remote Sensing (GEDI)

Light Detection and Ranging (LiDAR) is an active remote sensing technique that uses pulsed laser light to measure distances. In forestry, LiDAR is unique in its ability to provide direct 3D vertical measurements of forest structure, rather than inferring it from spectral or backscatter proxies.

Principles of GEDI Waveform LiDAR The Global Ecosystem Dynamics Investigation (GEDI) is a full-waveform LiDAR instrument mounted on the International Space Station (ISS). Unlike discrete-return LiDAR which records specific points, GEDI records the complete temporal profile (waveform) of the returned laser energy. This waveform represents the vertical distribution of intercepting surfaces from the canopy top to the ground.

GEDI Level 4A (L4A) Product For regional biomass mapping where wall-to-wall field plots are unavailable, GEDI data serves as a critical source of training data. The GEDI Level 4A (L4A) product provides footprint-level (approx. 25m diameter) estimates of Above-Ground Biomass Density (AGBD).

The L4A product is derived using a parametric framework that relates metrics extracted from the L2A waveform (such as Relative Height metrics, e.g., RH50, RH98) to field-measured biomass using extensive calibration plots from around the world. While GEDI L4A is a modeled product and not a direct measurement, it represents the highest quality "spaceborne plots" currently available. It allows researchers to upscale sparse samples into continuous maps by using GEDI footprints to train machine learning models based on wall-to-wall optical and radar imagery (Dubayah et al., 2022).

2.2.4 Sentinel – 1

The Sentinel-1 mission is a constellation of Earth observation satellites developed by the European Space Agency (ESA) under the Copernicus program. Designed for continuous, all-weather, day-and-night monitoring, the mission originally consisted of two satellites—Sentinel-1A (launched in 2014) and Sentinel-1B (launched in 2016)—operating in a sun-synchronous near-polar orbit. This dual-satellite configuration enabled a revisit time of 6–12 days depending on latitude, facilitating consistent global monitoring of land surface dynamics.

Sentinel-1 carries a C-band Synthetic Aperture Radar (SAR) instrument operating at 5.405 GHz (≈5.6 cm wavelength). For terrestrial applications, the Interferometric Wide (IW) swath mode is most commonly used. This study employs the Ground Range Detected (GRD) product, which provides multilooked amplitude data projected to ground range. Key technical characteristics relevant to biomass modeling include:

1. Polarization: The mission offers VV and VH dual-polarization data. VH polarization is especially useful for forest monitoring because it is more sensitive to volume scattering from vegetation structure compared to VV.

6. Spatial Resolution: The GRD IW data provide 10 m spatial resolution, allowing pixel-level integration with Sentinel-2 optical imagery.

Due to its relatively short wavelength, Sentinel-1's C-band interacts primarily with the upper canopy layer. Scattering models show that the radar signal is dominated by leaves, small branches, and twigs rather than the main woody components of the forest (Yu & Saatchi, 2016). The backscatter is also highly sensitive to moisture; rainfall or increased canopy wetness can significantly increase the return signal, introducing variability unrelated to biomass.

The main advantage of Sentinel-1 is its all-weather reliability. Unlike optical sensors, it can acquire imagery through cloud cover, haze, and smoke, ensuring consistent temporal coverage in humid tropical regions. However, the sensor's utility for Above-Ground Biomass (AGB) prediction is constrained by signal saturation. The C-band typically saturates between 60–100 Mg/ha, making it incapable of distinguishing high and very high biomass levels common in mature tropical forests (Tian et al., 2023). As a result, Sentinel-1 alone is insufficient for carbon stock estimation but remains highly valuable when integrated with longer-wavelength radar, optical data, or LiDAR observations.

2.2.5 Sentinel – 2

The Sentinel-2 mission consists of a constellation of two polar-orbiting satellites, Sentinel-2A and Sentinel-2B, developed by the European Space Agency (ESA). The primary instrument on board is the Multi-Spectral Instrument (MSI), a passive sensor designed to capture reflected solar radiation from the Earth's surface. The mission provides high-resolution optical imagery with a high revisit frequency (5 days at the equator with two satellites), which is critical for monitoring vegetation dynamics and land cover changes (Drusch et al., 2012).

The MSI sensor samples 13 spectral bands ranging from the visible and near-infrared (VNIR) to the short-wave infrared (SWIR).

1. Visible and NIR: The sensor captures standard visible bands (Blue, Green, Red) at 10-meter spatial resolution, which are essential for delineating vegetation extent.
7. Red-Edge Bands: A distinguishing feature of Sentinel-2 is its inclusion of three distinct bands in the "Red-Edge" spectrum (the transition zone between red absorption and NIR reflectance). These bands are particularly sensitive to chlorophyll content and have been shown to maintain sensitivity in high-density vegetation where standard red bands saturate (Delloye et al., 2018).

8. SWIR Bands: The Short-Wave Infrared (SWIR) bands (Band 11 and 12) are highly correlated with vegetation moisture content and structural complexity, serving as a strong proxy for biomass volume in healthy forests (Ma et al., 2020).

To quantify biomass from optical data, a diverse suite of Vegetation Indices (VIs) is utilized to isolate the vegetation signal from background noise and mitigate saturation effects. These can be categorized into three groups:

1. Greenness and Chlorophyll Indices (NDVI, GNDVI, NDRE) The Normalized Difference Vegetation Index (NDVI) is the standard metric for assessing vegetation health, calculated from the NIR and Red bands. However, NDVI is prone to saturation in dense canopies. To address this, the Green Normalized Difference Vegetation Index (GNDVI) uses the Green band instead of the Red, which is more sensitive to chlorophyll concentration. Furthermore, the Normalized Difference Red Edge (NDRE) utilizes the specific Red-Edge bands of Sentinel-2. NDRE is particularly effective for tropical biomass estimation as it remains sensitive to chlorophyll variations in high-density forests where traditional NDVI saturates (David et al., 2022).
9. Soil-Adjusted Indices (EVI, SAVI, MSAVI2) In areas with sparse canopy or open agroforestry systems, soil background reflectance can distort biomass estimates. The Enhanced Vegetation Index (EVI) incorporates blue band coefficients to correct for atmospheric aerosol scattering and soil background signals . Similarly, the Soil Adjusted Vegetation Index (SAVI) and Modified Soil Adjusted Vegetation Index 2 (MSAVI2) introduce adjustment factors to decouple the vegetation signal from the underlying soil brightness, ensuring more accurate predictions in heterogeneous landscapes (Adamu et al., 2021).
10. Moisture and Structure Indices (NDMI, NBR) Biomass volume is strongly correlated with canopy water content and woody structure. The Normalized Difference Moisture Index (NDMI) leverages the NIR and SWIR bands to detect vegetation water stress. Additionally, the Normalized Burn Ratio (NBR), while typically used for fire severity, is valuable for biomass mapping as it integrates SWIR reflectance, which is sensitive to the lignin and cellulose content of woody vegetation (trunks and branches) rather than just leaf pigment (Khan et al., 2020).

The primary strength of Sentinel-2 lies in its high spatial resolution (10 m), which allows for the precise characterization of the horizontal canopy distribution and species composition.

This high horizontal fidelity helps effectively delineate forest boundaries and remove non-forest confusion.

However, the use of optical sensors for biomass estimation is hindered by two significant limitations:

1. Optical Saturation: In complex tropical rainforests, the spectral signal primarily reflects the upper canopy closure. Once the canopy closes completely (typically around 100–150 Mg/ha), the relationship between reflectance and biomass becomes asymptotic, leading to significant underestimation in carbon-dense forests (Migolet et al., 2022).
11. Cloud Contamination: As a passive sensor relying on solar illumination, Sentinel-2 is heavily hampered by persistent cloud cover, which is prevalent in tropical regions like Indonesia, often resulting in data gaps that require aggressive temporal filtering (Slagter et al., 2023).

2.2.6 ALOS PALSAR – 2

The Advanced Land Observing Satellite-2 (ALOS-2), or DAICHI-2, is a Japanese Earth observation mission launched by JAXA on May 24, 2014 (JAXA, 2015). It is the successor to ALOS (2006–2011) and is optimized for all-weather, day-night imaging using L-band Synthetic Aperture Radar (SAR). Its mission supports disaster monitoring, forest management, agriculture, land deformation, and assessment of biomass and carbon stocks in tropical forests, where persistent cloud cover limits optical remote sensing (Rosenqvist et al., 2014).

ALOS-2 carries the Phased Array L-band Synthetic Aperture Radar-2 (PALSAR-2), operating at an L-band wavelength of approximately 23.6–24 cm (center frequency ~1.236 GHz). Compared to shorter wavelength systems (e.g., C-band), L-band penetrates deeper into forest canopies and interacts with trunks and larger branches, making it valuable for tropical forest structural mapping. PALSAR-2 supports multiple polarization modes—HH, HV, VH, and VV (dual-pol or quad-pol depending on the acquisition mode). Its spatial resolutions range from 1–3 m (Spotlight) to 6 m (High-Sensitive Stripmap) and 100 m (ScanSAR), with swath widths up to 350 km and a 14-day repeat cycle.

L-band microwaves are sensitive to volume scattering from woody components, especially in HV polarization, which is strongly correlated with forest structural complexity (Tello et al., 2018). HV backscatter correlates with aboveground biomass (AGB) up to ~200–300 Mg/ha, and in some forest types up to ~400–500 Mg/ha before saturation This substantially

exceeds the saturation threshold of optical vegetation indices and C-band SAR (typically 60–100 Mg/ha).

However, ALOS-2 acquisitions depend on JAXA's tasking priorities, leading to 14–28 day effective revisit times over the tropics, limiting multi-temporal analysis compared to systems like Sentinel-1. Additionally, speckle noise, topographic distortions, and incidence-angle variability require extensive preprocessing before fusion with optical or LiDAR-derived datasets.

2.2.7 GEDI

The Global Ecosystem Dynamics Investigation (GEDI) is a NASA mission deployed to the International Space Station in December 2018, designed to deliver the first global spaceborne lidar observations of forest canopy height and vertical structure across latitudes from 51.6° north to 51.6° south. Operating from the ISS, GEDI employs three lasers that generate eight distinct ground tracks, allowing detailed sampling of forest vertical profiles. These measurements support accurate estimation of aboveground biomass (AGB), which is particularly important for modeling carbon stocks in tropical regions. The ISS orbital configuration provides frequent and dense observations over the tropics, including Indonesia, throughout GEDI's primary operational phase, which concluded in 2023, while the archived datasets continue to be used for integrated and fusion-based analyses.

GEDI employs full-waveform LiDAR with a footprint size of approximately 25 m diameter, capturing the complete return signal from laser pulses (1064 nm wavelength, 242 Hz pulse rate) to measure canopy vertical structure. The full-waveform approach records the time-of-flight and intensity of scattered photons, allowing decomposition into ground, canopy layers, and gaps. Key vertical structure metrics include Relative Height (RH) metrics (e.g., RH98 for top-of-canopy height, RH50 for mean canopy height), which quantify canopy layering and volume for biomass prediction.

GEDI Level 2A (L2A) products provide geolocated waveforms and processed metrics like RH percentiles, canopy cover, and plant area index (PAI) at 25 m resolution. Level 4A (L4A) products deliver calibrated AGB estimates (Mg/ha) and AGB density (Mg C/ha), trained on field data from tropical forests up to ~250 Mg/ha, with uncertainty flags for quality control. These products are available via NASA's Earthdata portal, supporting multi-sensor fusion with Sentinel-2 and ALOS PALSAR-2 for wall-to-wall mapping.

GEDI offers direct, physics-based measurements of forest structural attributes like canopy height and layering, improving AGB accuracy in complex tropical canopies where optical/SAR sensors saturate or lack vertical sensitivity.

Sparse footprint sampling (e.g., ~1 km track spacing, 60 m between tracks) covers only ~10% of forests directly, requiring wall-to-wall prediction models (e.g., regression with optical/radar data) for continuous maps. ISS orbit also limits revisit frequency and introduces geolocation errors in dense tropics.

2.3 Multi-Sensor Data Fusion

As forest ecosystems are complex, three-dimensional structures, no single remote sensing sensor is capable of capturing all physical attributes required for accurate biomass estimation. Multi-sensor data fusion has therefore emerged as a standard methodological framework in modern biometry. This section discusses the theoretical basis for sensor synergy and reviews recent applications involving GEDI data integration.

2.3.1 The Rationale for Synergistic Fusion

Data fusion refers to the integration of information from multiple data sources to produce outputs that are both more accurate and more informative than those derived from any single source alone (Castanedo, 2013). Within forest biomass mapping applications, data fusion is used to overcome the inherent constraints associated with individual sensing technologies.

Complementarity of Horizontal and Vertical Information The primary advantage of fusing optical and radar data lies in the integration of "horizontal" and "vertical" spectral information:

1. Horizontal Structure (Optical Data): Optical sensors, such as Sentinel-2, provide high-resolution information regarding the horizontal canopy distribution. They excel at delineating vegetation extent, species composition (via spectral signatures), and canopy closure. However, as noted in Section 2.2.1, they saturate quickly and fail to capture volume.
12. Vertical Structure (Radar/SAR): Long-wavelength radar, such as L-Band SAR, interacts with the vertical profile of the forest. It provides information on trunk volume, branch structure, and moisture content. However, SAR data is often plagued by speckle noise and geometric distortions caused by topography.

The Synergistic Effect By combining these datasets, researchers can exploit a synergistic effect. The optical data helps to characterize the canopy surface and remove non-forest

confusion, while the radar data adds depth and volumetric information. This combination significantly delays the saturation point, allowing for more accurate regression in high-biomass tropical forests compared to single-sensor approaches (Antunes et al., 2024).

2.3.2 Review of GEDI Fusion Studies

The availability of GEDI data has shifted the paradigm of fusion from simple Optical-Radar combinations to a three-tier approach: using LiDAR for sampling and Optical/Radar for spatial extrapolation.

Recent studies have demonstrated the efficacy of this approach in tropical environments:

- GEDI + Sentinel-1/2: Research by (Liu et al., 2025) demonstrated that fusing GEDI waveform metrics with Landsat and Sentinel data allowed for the creation of global forest height maps. The study found that while GEDI provided accurate vertical anchor points, the optical time-series data was crucial for capturing phenological variations and filling gaps between GEDI tracks.
- GEDI + L-Band SAR: A study by (Musthafa & Singh, 2022) highlighted that adding L-Band SAR to LiDAR samples significantly improves biomass estimation in dense tropical forests compared to using C-Band or optical data alone. The L-Band backscatter correlates strongly with the height metrics derived from LiDAR, stabilizing the model predictions in areas where optical indices saturate.
- Multi-Layer Perceptron Integration: More recently, Wang et al. (2024) successfully utilized GEDI samples to train Multi-Layer Perceptron models using fused satellite imagery. However, a critical methodological limitation of their study was the absence of a wrapper-based feature selection strategy. By feeding the model high-dimensional multi-sensor data without systematic pruning, their approach remains vulnerable to the 'curse of dimensionality' and feature redundancy, which can degrade the convergence and generalization of neural networks. This thesis specifically addresses this gap by implementing Recursive Feature Elimination (RFE). unlike the static input sets used in previous studies, RFE dynamically identifies the optimal subset of features that contributes most to model accuracy, removing noise while preserving physical signal.

These studies collectively validate the methodology proposed in this thesis: utilizing GEDI L4A as the reference training data to calibrate a fusion model based on the horizontal fidelity of Sentinel-2 and the vertical penetration of ALOS PALSAR-2.

2.4 Feature Selection

Feature selection is the process of identifying the most relevant predictor variables (features) that contribute significantly to model performance. In multi-sensor remote sensing applications—where datasets may include dozens of spectral bands, SAR backscatter values, vegetation indices, and ancillary variables—feature selection is essential for reducing redundancy, minimizing noise, and improving prediction accuracy. It also decreases computational cost, particularly for algorithms such as SVR and MLP that are sensitive to high-dimensional inputs.

2.4.1 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a wrapper-based feature selection technique designed to identify the optimal subset of predictors that maximizes model performance. Unlike filter methods that rank features based on statistical metrics (e.g., correlation) independent of the model, RFE operates by repeatedly training a specific machine learning estimator—in this study, the Random Forest Regressor—and pruning the least significant variables based on their model-derived importance weights

The RFE process is iterative and follows a backward elimination strategy:

1. Initialization: The algorithm begins by training the base estimator (Random Forest) on the full set of available features.
13. Ranking: Feature importance scores are extracted from the trained model. For Random Forest, this is typically the Mean Decrease Impurity (MDI).
14. Elimination: The feature(s) with the lowest importance score are removed from the dataset.
15. Iteration: Steps 1 through 3 are repeated on the progressively smaller feature subsets until a specified number of features remains.

This iterative process ensures that the algorithm accounts for the changing relative importance of features as redundancy is removed.

The primary advantage of RFE in remote sensing applications is its ability to consider nonlinear relationships and feature interactions. By utilizing the actual model (Random Forest)

to determine importance, RFE captures complex synergies—such as the interaction between L-band backscatter and optical Red-Edge indices—that simpler filter methods might miss.

However, RFE is computationally expensive because it requires retraining the model at every step of the elimination process. Its effectiveness is also heavily dependent on the stability of the base estimator; if the underlying model is unstable (e.g., highly sensitive to small data changes), the resulting feature ranking may vary significantly. Despite these computational costs, RFE provides a robust method for reducing the high dimensionality of multi-sensor datasets while preserving predictive accuracy.

2.4.2 Mutual Information

Mutual Information (MI) is a non-parametric statistical measure that quantifies the amount of information obtained about one random variable by observing another. Unlike standard correlation coefficients (such as Pearson's r), which assess only linear relationships, MI measures the general dependence between variables. Formally, it calculates the reduction in uncertainty (entropy) of a target variable (e.g., Biomass) given the knowledge of a predictor variable (e.g., Radar Backscatter) (Papaioannou et al., 2025).

Mathematically, if X and Y are two random variables, the mutual information $I(X;Y)$ is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2.4}$$

Where $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions. If X and Y are independent, $p(x,y) = p(x)p(y)$, resulting in an MI of zero.

Mutual Information is particularly advantageous for multi-sensor biomass mapping due to three key characteristics:

1. Nonlinear Sensitivity: The relationship between satellite signals and biomass is frequently nonlinear, particularly due to signal saturation in optical and C-band radar data. While correlation-based approaches might underestimate the strength of these relationships (e.g., the asymptotic curve of NDVI vs. Biomass), MI captures the full dependency structure, identifying predictors that have strong predictive power even if the relationship is complex.

16. Distribution Robustness: Remote sensing datasets often exhibit non-Gaussian or mixed distributions (e.g., multi-modal distributions in heterogeneous landscapes like Yogyakarta). MI is robust to these irregularities and does not assume normality, unlike parametric tests such as ANOVA.
17. High-Dimensional Compatibility: MI is effective for filtering features in high-dimensional datasets, allowing researchers to evaluate the relevance of dozens of spectral indices and texture metrics without the risk of multicollinearity distorting the metric itself.

Despite its theoretical strengths, MI presents certain practical challenges:

1. Sample Size Sensitivity: The accurate estimation of the joint probability distribution $p(x, y)$ requires a substantial number of samples. In data-scarce environments, MI estimates can be biased or noisy.
2. Interpretability: Unlike correlation, which provides a bounded value $(-1\ to\ +1)$ indicating both strength and direction (positive or negative relationship), MI provides a non-negative value measured in "bits" or "nats." This makes it harder to intuitively compare the magnitude of relationships or determine the directionality of the effect without additional analysis.

2.4.3 SelectKBest with ANOVA F-test

SelectKBest is a univariate feature selection method that evaluates each feature individually to determine its relationship with the target variable. In the context of regression tasks (such as biomass estimation), this technique typically utilizes the ANOVA (Analysis of Variance) F-test as the scoring function (Abdumalikov et al., 2024; Pritalia, 2022).

The F-test assesses the strength of the linear dependency between a specific predictor (e.g., a single spectral band) and the target variable (AGB). Mathematically, it calculates the F-value, which represents the ratio of the variance explained by the feature to the unexplained variance (error). Features are then ranked based on these F-scores, and the top k features with the highest scores are retained for model training .

The primary advantage of the SelectKBest approach is its computational efficiency. Unlike wrapper methods such as Recursive Feature Elimination (RFE) that require training a model iteratively, the ANOVA F-test is a simple statistical calculation that can be performed almost instantaneously, even on very large datasets. This makes it an ideal strategy for the initial screening of high-dimensional feature sets, allowing researchers to quickly discard

irrelevant variables and reduce redundant spectral bands or indices before applying more computationally intensive algorithms.

Despite its speed, the ANOVA F-test has significant limitations in complex remote sensing applications. First, it strictly detects linear relationships; it may discard features that have strong but nonlinear correlations with biomass (e.g., radar backscatter saturation curves). Second, as a univariate method, it evaluates features independently, ignoring feature interactions and redundancy; this limitation can lead to the loss of important multi-feature information that enhances predictive performance. Finally, because it relies on simple variance comparisons for each individual feature, it can select features that are statistically significant yet practically suboptimal when interactions are considered (Cheng, 2024; Pudjihartono et al., 2022).

2.4.4 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised dimensionality reduction technique widely used in remote sensing to transform high-dimensional datasets into a smaller set of variables while retaining the majority of the original information. It achieves this by transforming the original correlated features (e.g., spectral bands) into a new set of uncorrelated, orthogonal variables known as Principal Components (PCs) (Lu & Weng, 2007).

The mechanism of PCA involves four key mathematical steps:

1. Covariance Computation: The algorithm first computes the covariance matrix of the standardized input features to understand how variables vary with respect to one another.
18. Eigen-decomposition: It extracts the eigenvalues and eigenvectors from the covariance matrix. The eigenvectors determine the direction of the new feature space, while the eigenvalues determine the magnitude (variance) explained by each direction.
19. Projection: The original data is projected into this new coordinate system. The first Principal Component (PC1) is aligned with the direction of maximum variance in the data. The second component (PC2) captures the second highest variance while remaining orthogonal (perpendicular) to PC1, and so on.
20. Component Selection: Finally, the top k components that explain a cumulative variance threshold (e.g., 95%) are selected, while the remaining components (often representing noise) are discarded.

In the context of multi-sensor biomass mapping, PCA offers several distinct advantages. Primarily, it effectively reduces multicollinearity. Satellite datasets often contain highly

correlated variables (e.g., NDVI is highly correlated with EVI, and Sentinel-2 Band 2 is correlated with Band 3). PCA consolidates these into independent components, stabilizing parametric models like Linear Regression. Additionally, it improves computational efficiency by reducing the number of input variables the model must process, and it is capable of capturing the underlying latent structure of complex multispectral and SAR datasets.

Despite its utility, PCA has significant limitations for biophysical parameter retrieval. The most critical drawback is the loss of interpretability. The resulting Principal Components are linear combinations of the original spectral bands and lack direct physical meaning (e.g., one cannot say "PC1 represents chlorophyll absorption"). Furthermore, PCA relies on the assumption that variance equals information. In biomass estimation, important biological signals might exhibit low variance compared to atmospheric or topographic noise; PCA risks discarding these subtle but critical features. Finally, as a linear transformation technique, PCA may fail to unfold the complex non-linear manifolds inherent in tropical forest data (Hoffmann, n.d.).

2.5 Machine Learning in Remote Sensing

The retrieval of biophysical parameters such as Above-Ground Biomass (AGB) from satellite data is mathematically posed as an inverse problem. Traditional parametric methods (e.g., linear regression) often fail to resolve the complex, non-linear relationships between satellite signals and forest structure. Consequently, Machine Learning (ML) algorithms have become the standard for developing robust predictive models. This section reviews the three distinct algorithms evaluated in this study.

2.5.1 Random Forest

Random Forest (RF) is an ensemble learning method that constructs a multitude of decision trees and averages their predictions. In remote sensing, RF is widely considered one of the most robust algorithms due to its ability to handle high-dimensional data and resist overfitting (Antunes et al., 2024; Zhang et al., 2022).

Working Principle RF employs a technique known as "bootstrap aggregating" or *bagging*. The algorithm creates multiple subsets of the training data through random sampling with replacement. A decision tree is grown for each subset, and at each node split, only a random subset of features is considered. For regression tasks (like biomass estimation), the final prediction is the average of the outputs of all individual trees:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \quad (2.4)$$

Equation 2.4 illustrates the averaging mechanism central to the Random Forest regression model. In this equation, the final predicted value $\hat{y}$ is derived by calculating the arithmetic mean of the outputs from all individual decision trees in the ensemble. The variable N represents the total number of trees generated during the training phase, while $T_i(x)$ denotes the specific prediction made by the i tree for the input vector x. This averaging process is the mathematical basis for the algorithm's stability, as it mitigates the variance and overfitting risks often associated with single decision trees.

Strengths in Remote Sensing The primary advantage of RF is its stability. By averaging the results of noisy, unbiased trees, it reduces the variance of the model without increasing the bias. It is particularly effective at handling outliers and noise in satellite data, making it a reliable benchmark for biomass mapping studies (Belgiu & Drăgu, 2016).

2.5.2 Multi-Layer Percepton

The Multi-Layer Perceptron (MLP) is a class of feedforward Artificial Neural Network (ANN) that serves as a powerful tool for modeling complex, non-linear relationships in remote sensing data. Unlike traditional statistical models that assume linear dependencies, the MLP is designed to approximate continuous functions by learning from data through a network of interconnected processing units called neurons.

Structurally, an MLP consists of at least three distinct layers: an input layer, one or more hidden layers, and an output layer (Ferdinandus Edwin Penalun et al., 2024). In the context of biomass estimation, the input layer receives the feature vector (e.g., spectral bands from Sentinel-2, backscatter coefficients from Sentinel-1). These signals are then propagated forward through the hidden layers. Each neuron in a hidden layer performs a specific computation: it calculates the weighted sum of inputs from the previous layer, adds a bias term, and passes the result through a non-linear activation function.

Mathematically, The output y_j of a neuron j is expressed as:

$$y_j = \varphi(\sum_{i=1}^{n} w_{ij}\, x_i + b_j) \quad (2.5)$$

Where:

- x_i represents the input signals.
- w_{ij} denotes the synaptic weights connecting neuron i to neuron j.
- b_j is the bias term.

- φ is the non-linear activation function (commonly the Rectified Linear Unit or ReLU, or the Sigmoid function).

The primary strength of the MLP lies in its learning mechanism, known as backpropagation. As shown in Equation (2.5), each neuron computes its output by applying a non-linear activation function to the weighted sum of the input features. During training, the network makes a prediction, compares it with the actual target value (e.g., GEDI-derived biomass), and calculates the error using a loss function such as Mean SquaredError. The backpropagation algorithm then propagates this error backward through the network, iteratively adjusting the weights w_{ij} to minimize the prediction error (Singh et al., 2022).

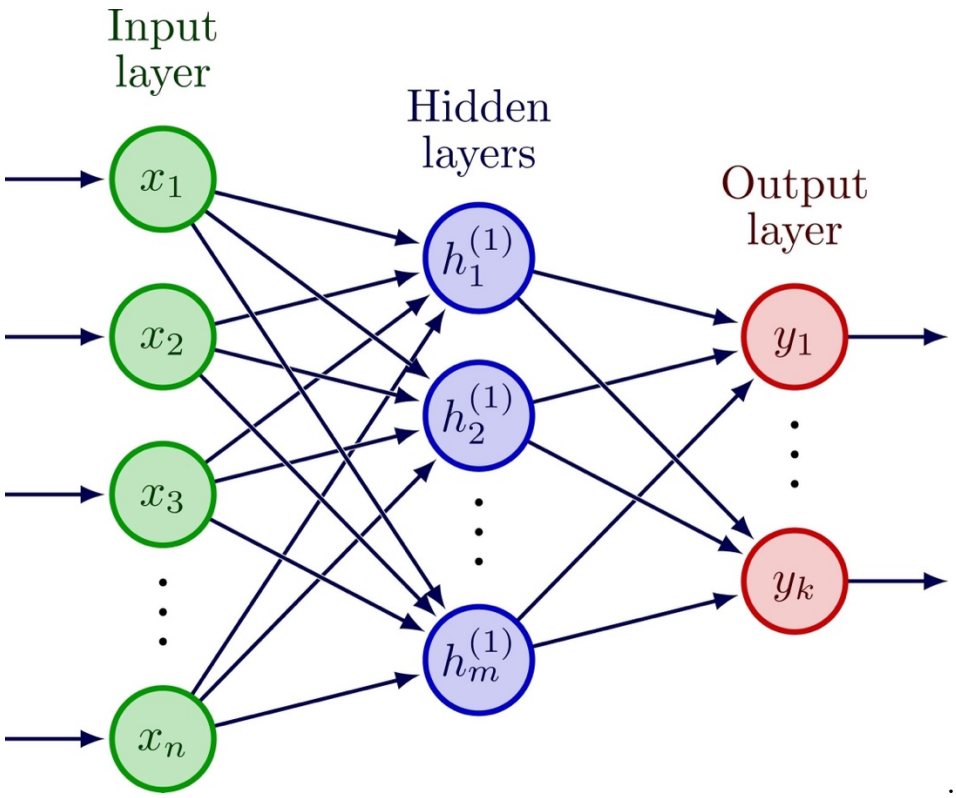


Figure 2.1 Architecture of a feedforward neural network with an input layer, one hidden layer, and an output layer.

The primary strength of the MLP lies in its learning mechanism, known as backpropagation, as illustrated in Figure 2.1. During the training phase, the network makes a prediction, compares it to the actual value (e.g., GEDI-derived biomass), and calculates the error using a loss function (such as Mean Squared Error). The algorithm then propagates this error backward through the network, iteratively adjusting the weights $w_{ij}$to minimize prediction error (Singh et al., 2022). This allows the MLP to automatically construct high-level

abstract features from raw sensor data, capturing complex interactions between optical and radar signals that are often difficult to define manually.

However, a critical requirement for effective MLP implementation is data scaling. Unlike decision tree-based algorithms (e.g., Random Forest), which are invariant to the scale of input features, MLPs are highly sensitive to feature magnitudes. Inputs with large ranges (e.g., elevation in meters) can dominate the gradient descent process over inputs with small ranges (e.g., reflectance values between 0 and 1), leading to poor convergence. Therefore, preprocessing steps such as Min-Max normalization or Z-score standardization are essential to ensure all features contribute equally to the learning process.

2.5.3 Support Vector Regression

Support Vector Regression (SVR) is an extension of the Support Vector Machine (SVM) algorithm, originally developed by Vapnik for classification tasks (Vapnik, 1995). While SVM fits a hyperplane to separate classes, SVR adapts this principle to regression problems by fitting a hyperplane that best predicts continuous variables. In the context of remote sensing, SVR is particularly valued for its robust capacity to model complex, nonlinear relationships between satellite spectral signatures and biophysical parameters (such as biomass) without making strong assumptions about the underlying data distribution.

The fundamental goal of SVR differs from traditional regression methods like Multiple Linear Regression (MLR), which aim to minimize the sum of squared errors between the predicted and actual values. Instead, SVR attempts to fit a function $f(x)$ that deviates from the actual target values y_i by a value no greater than a specified margin of tolerance, denoted as ε (epsilon), while simultaneously keeping the model as "flat" or smooth as possible (Rodríguez-Pérez & Bajorath, 2022).

This approach creates an ε-insensitive zone (or "tube") around the regression function. Errors falling within this tube are ignored, meaning they do not contribute to the loss function. Penalties are only assigned to data points that fall outside this margin. These boundary-defining points are known as "support vectors," and they are the only data points that influence the final regression model.

The linear regression function in the feature space can be expressed as:

$$f(x) = w^T x + b \tag{2.6}$$

Where w is the weight vector determining the orientation of the hyperplane, and b is the bias term.

To balance prediction accuracy (minimizing error) with model generalization (maximizing flatness/smoothness), SVR solves a convex optimization problem. The objective is to minimize the norm of the weights $||w||^2$ while penalizing errors that exceed the ε threshold. The primal optimization problem is formulated as:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}|w|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2.7}$$

Subject to the constraints:

$$\begin{aligned} y_i - f(x_i) &\leq \epsilon + \xi_i \\ f(x_i) - y_i &\leq \epsilon + \xi_i^* \\ \xi_i,\, \xi_i^* &\geq 0 \end{aligned} \tag{2.8}$$

Where:

- C (Penalty Parameter): Controls the trade-off between the flatness of the model and the tolerance for deviations larger than ϵ. A high C penalizes errors heavily (risking overfitting), while a low C allows for more errors (smoother model).
- ϵ (Epsilon): Defines the width of the insensitive zone (tube) within which errors are ignored.
- ξ_i, ξ_i^* (Slack Variables): Represent the magnitude of the error for data points that fall outside the ϵ-tube (above and below, respectively).

To handle the non-linear relationships inherent in multi-sensor biomass estimation (e.g., the complex interaction between radar backscatter and canopy structure), SVR utilizes the "kernel trick." This involves mapping the input vectors into a higher-dimensional feature space where linear regression can be performed.

Common kernel functions include the Linear, Polynomial, and Radial Basis Function (RBF) kernels. The RBF kernel is widely used in remote sensing due to its effectiveness in handling localized variations. It is mathematically defined as:

$$K(x_i, x_j) = \exp(-\gamma|x_i - x_j|^2) \tag{2.9}$$

Where γ (gamma) defines the influence of a single training example; low values mean 'far' and high values mean 'close'.

SVR offers several distinct advantages for forestry applications:

1. Robustness to High Dimensionality: It performs well even when the number of input features is large relative to the number of samples, making it suitable for multi-sensor fusion (Sentinel-1, Sentinel-2, ALOS PALSAR).
21. Outlier Resistance: By relying on the ϵ-insensitive loss function, SVR is less sensitive to noisy data points or outliers compared to algorithms using squared-error loss (like MLR or standard Neural Networks).
22. Non-Linear Modeling: Through the use of kernels, SVR can effectively model the asymptotic saturation behavior often observed between satellite signals and high-biomass forests.

Despite its strengths, SVR presents specific challenges:

1. Sensitivity to Feature Scaling: SVR is highly sensitive to the scale of input variables. Features with large ranges (e.g., Elevation in meters) can dominate those with small ranges (e.g., Reflectance < 1). Therefore, data standardization (e.g., Z-score normalization) is a strict prerequisite for SVR, similar to Multi-Layer Perceptrons.
23. Computational Cost: The performance of SVR relies heavily on the optimal selection of hyperparameters (, , and). Tuning these parameters (e.g., via Grid Search) is computationally expensive and time-consuming.
24. Scalability: Training time scales cubically with the number of samples, which can be a bottleneck when using large datasets such as region-wide GEDI footprints.

2.5.4 Multi Linear Regression

Multiple Linear Regression (MLR) is a fundamental statistical method used to model the linear relationship between a single continuous dependent variable (in this case, Above-Ground Biomass or AGB) and two or more independent predictor variables (such as optical spectral bands, SAR backscatter coefficients, and LiDAR metrics). Historically, MLR has served as the baseline approach in forest biomass and ecological modeling due to its simplicity and well-understood theoretical properties. It assumes that the target variable can be expressed as a weighted sum of the input features plus an error term (Opelele et al., 2021).

The general form of the multiple linear regression model is expressed mathematically as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (2.10)$$

Where:

- y: Represents the dependent variable, specifically the Above-Ground Biomass (AGB) in $Mg \cdot ha^{-1}$.
- $\boldsymbol{x_1, x_2, \ldots, x_n}$: Represent the predictor variables derived from satellite sensors (e.g., NDVI, Sentinel-1 VV backscatter, Elevation).
- $\boldsymbol{\beta_0}$: Is the y-intercept (constant term).
- $\boldsymbol{\beta_1, \ldots, \beta_n}$: Are the regression coefficients (slopes) that quantify the change in y for a one-unit change in the corresponding predictor x, holding all other variables constant.
- $\boldsymbol{\epsilon}$: Is the error term (residual), representing the variance in y that is not explained by the linear model.

The regression coefficients (β) are typically estimated using the Ordinary Least Squares (OLS) method. The objective of OLS is to find the set of coefficients that minimizes the sum of the squared differences (residuals) between the observed values (y_i) and the predicted values ($\widehat{y_i}$). The cost function to be minimized is:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \quad (2.10)$$

By minimizing this quantity, the algorithm fits a hyperplane through the multi-dimensional feature space that lies "closest" to all data points.

For MLR estimates to be statistically valid and unbiased, several key assumptions must be met:

1. Linearity: The relationship between the independent and dependent variables is linear.
25. Independence of Errors: The residuals are not correlated with each other (no autocorrelation).
26. Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables.
27. Normality of Residuals: The error terms follow a normal distribution.
28. No Multicollinearity: The independent variables are not highly correlated with each other.

In the context of remote sensing, these assumptions are frequently violated. For instance, biomass data often exhibits spatial autocorrelation (violating independence), and spectral bands from the same sensor (e.g., Red and Green bands) are often highly correlated (violating the no multicollinearity assumption).

Despite the rise of advanced machine learning, MLR remains relevant due to several strengths:

1. Interpretability: It provides direct insight into the relationship between variables. The magnitude and sign of the coefficients (β) explicitly reveal how each sensor input contributes to the biomass estimate.
29. Computational Efficiency: OLS estimation is computationally inexpensive and fast, even for large datasets.
30. Simplicity: It serves as an excellent baseline benchmark for evaluating the performance gains of more complex models.

However, MLR has significant limitations when applied to complex biological systems like tropical forests:

- Non-Linearity: As noted in recent studies, the relationship between satellite signals (e.g., reflectance) and forest structure is inherently non-linear2. MLR forces a linear fit, which often fails to capture the complex saturation dynamics.
- Multicollinearity: Remote sensing datasets often contain highly correlated features (e.g., NDVI and EVI), which can lead to unstable coefficient estimates in MLR.
- Sensitivity to Outliers: As a parametric method minimizing squared errors, MLR is highly sensitive to outliers, which can skew the regression line significantly.
- Underestimation of Extremes: MLR typically struggles at the high end of the biomass distribution, leading to systematic underestimation of carbon-dense primary forests, as confirmed by the comparison results in this study where MLR yielded the lowest R^2 (0.2976)3.

CHAPTER III
RESEARCH METHODS

This chapter outlines the data and the procedures used to achieve the research objectives. It begins with a description of the study area, followed by a detailed list of the satellite and ground-truth datasets. Finally, it presents the complete methodological workflow, from data pre-processing and feature engineering to the final machine learning model training and validation.

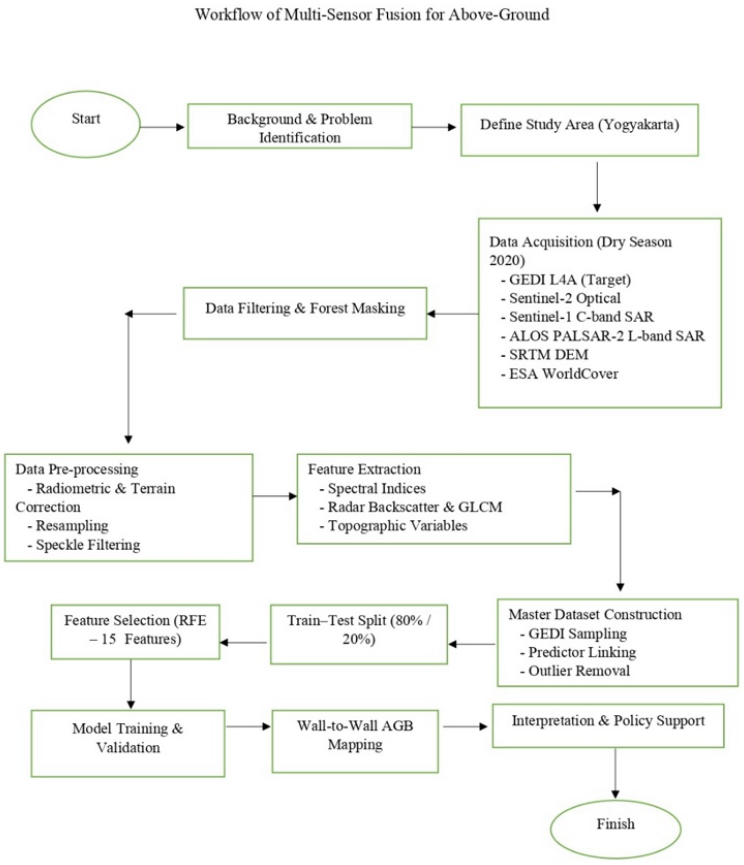


Figure 3.1 Workflow for Predicting mapping

3.1 Study Area

The research was conducted in the Special Region of Yogyakarta (Daerah Istimewa Yogyakarta - DIY), a province located on the southern coast of Java, Indonesia. Geographically, the study area is situated between the coordinates 7°33′–8°12′ South Latitude and 110°00′–110°50′ East Longitude. The region encompasses a total land area of approximately 3,185.80 km² and is administratively divided into four regencies (Sleman, Bantul, Kulon Progo, and Gunung Kidul) and one city (Yogyakarta). The spatial orientation and administrative boundaries of the study location are presented in Figure 3.1.

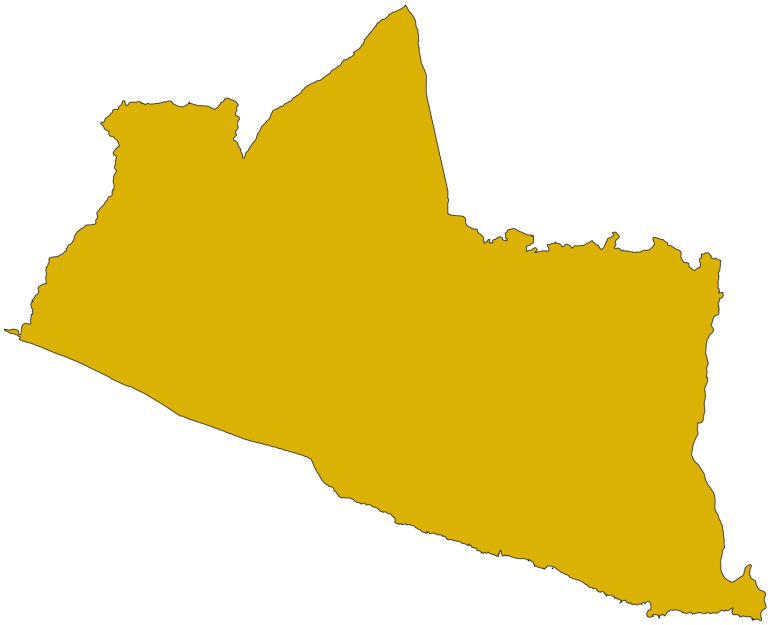


Figure 3.2 Map of the study area showing the administrative boundaries of the Special Region of Yogyakarta and the distribution of forest cover types.

The physiography of Yogyakarta is characterized by distinct heterogeneity, which presents a significant challenge and opportunity for evaluating remote sensing models. The landscape is dominated by the volcanic axis in the north, centered on Mount Merapi, an active stratovolcano reaching an elevation of over 2,900 meters above sea level. This northern zone is characterized by steep slopes and dense tropical montane forests. In contrast, the western and southeastern regions—specifically the Menoreh Hills in Kulon Progo and the Gunung Sewu range in Gunung Kidul—are dominated by Karst (limestone) topography. These Karst regions exhibit complex terrain with irregular hills and ridges, which pose specific challenges for Synthetic Aperture Radar (SAR) geometric correction due to layover and shadowing effects.

Climatologically, the study area experiences a tropical monsoon climate (Type Am according to the Köppen-Geiger classification). The region is marked by two distinct seasons: a dry season typically spanning from June to September and a wet season from October to March. The average annual precipitation ranges between 2,000 mm and 3,000 mm, with temperatures averaging 26 °C to 28 °C in the lowlands, decreasing with altitude in the northern highlands. This seasonal variability necessitates the careful temporal filtering of satellite imagery to minimize atmospheric noise and moisture-induced dielectric fluctuations in radar signals.

In terms of land cover, the forest ecosystems in Yogyakarta are diverse. They range from the conservation forests of Mount Merapi National Park to extensive monoculture plantations (primarily Teak and Pine) managed by Perhutani in the Dlingo and Playen zones. Additionally, a significant portion of the biomass is contained within "Hutan Rakyat" (Community Forests), which are complex agroforestry systems combining timber species (e.g., Falcataria moluccana, Tectona grandis) with agricultural crops. This mixture of dense primary forest, plantation forest, and sparse agroforestry provides a wide range of biomass values, making the region an ideal site for testing the sensitivity of Machine Learning and Deep Learning models to varying carbon stock densities.

3.2 Data Acquisition

The integrity of multi-sensor fusion relies heavily on the temporal consistency of the input datasets. To minimize errors associated with phenological changes (e.g., leaf shedding in deciduous species) and atmospheric variability (e.g., cloud cover and soil moisture fluctuations), this study specifically focuses on the 2020 Dry Season, defined here as the period from June 1, 2020, to October 31, 2020. This timeframe coincides with the minimum cloud cover in the Java region, ensuring the highest quality optical observations and stable dielectric properties for radar backscatter.

All satellite data processing and acquisition were executed using the Google Earth Engine (GEE) cloud computing platform, which allows for the efficient handling of petabyte-scale datasets. A detailed summary of the acquired datasets and their specific product levels is provided in Table 3.1.

Table 3.1 Summary of satellite sensor specifications, temporal coverage, and spatial resolution used in the study.

Data Type	Product	Source	Key Features Used	Purpose
Target Data	GEDI L4A	NASA Earthdata(*GEDI L4A Footprint Level Aboveground Biomass Density, Version 2.1 - Earthdata Search*, n.d.)	agbd (Aboveground Biomass)	Ground-truth (Y)
Optical	Sentinel-2 L2A	Google Earth Engine (*Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A (SR))*	10 Bands + 8 Indices (NDVI, EVI, etc.)	Predictor (X)
C-Band Radar	Sentinel-1 GRD	Google Earth Engine (*Sentinel-1 SAR GRD: C-Band Synthetic Aperture Radar Ground Range Detected, Log Scaling)*	VV, VH Backscatter + 7 GLCM Textures	Predictor (X)
L-Band Radar	ALOS PALSAR	Google Earth Engine (*PALSAR-2 ScanSAR Level 2.2.*)	HH, HV Backscatter + Ratio	Predictor (X)
Topography	SRTM 30m	Google Earth Engine (*NASA SRTM Digital Elevation 30m)*	Elevation, Slope, Aspect	Predictor (X)
Filter	ESA WorldCover	Google Earth Engine (*ESA WorldCover 10m V100)*	Land Cover Class (Map)	Filtering Data

3.2.1 GEDI Level 4A Biomass Data (Reference Data)

The primary reference dataset for this study is derived from the Global Ecosystem Dynamics Investigation (GEDI), a spaceborne LiDAR instrument mounted on the International Space Station (ISS). Specifically, the Level 4A (L4A) product was utilized, which provides footprint-level estimates of Above-Ground Biomass Density (AGBD) in $Mg \cdot ha^{-1}$

To ensure high-quality training data, strict filtering criteria were applied to the GEDI granules acquired during the 2020 dry season. Only footprints satisfying the following conditions were retained:

1. Quality Flag: l4_quality_flag = 1 (Indicating valid waveform processing).
31. Degrade Flag: degrade_flag = 0 (Ensuring the pointing stability of the ISS was optimal).
32. Sensitivity: Beam sensitivity > 0.9 (To ensure the laser energy was sufficient to penetrate the canopy and reach the ground).

These filtering steps resulted in a set of high-quality discrete sampling points distributed across the study area. While GEDI L4A products are modeled estimates rather than direct field measurements, they provide calibrated, high-fidelity reference data suitable for training Machine Learning models in the absence of wall-to-wall National Forest Inventory plots. Consequently, this study utilizes these filtered footprints as the target variable for biomass prediction.. The spatial distribution of these valid GEDI footprints over Yogyakarta is visualized in Figure 3.2.

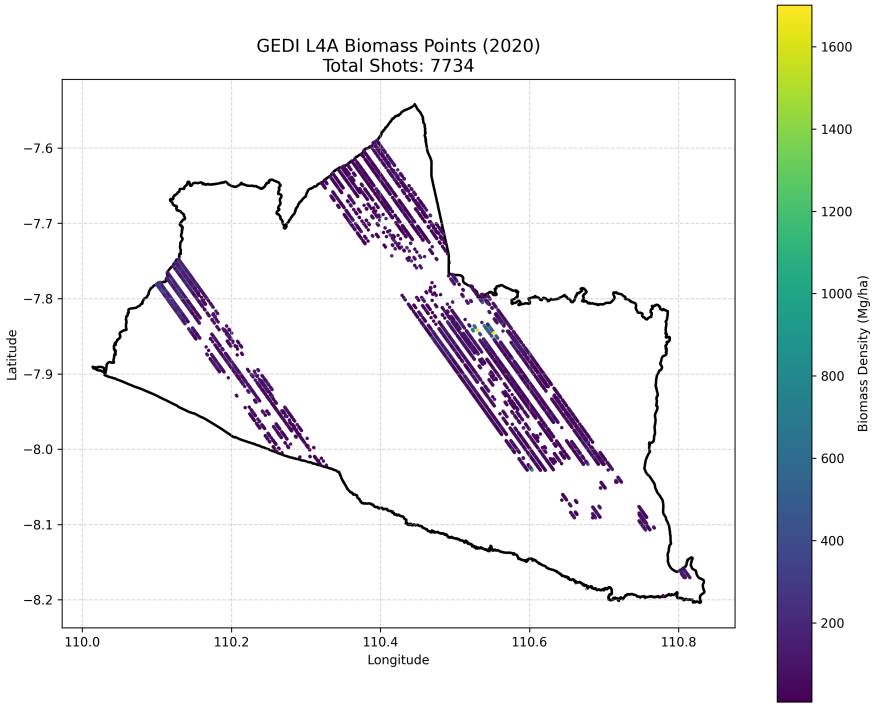


Figure 3.3 Spatial distribution of valid GEDI L4A footprints across the Special Region of Yogyakarta after quality filtering.

Note: This data is derived from the NASA GEDI mission (Level 4A product), acquired during the 2020 observation cycle. Although GEDI is a global product, this subset has been spatially filtered to the specific administrative boundaries of the study area.

3.2.2 Sentinel-2 MSI Imagery (Optical)

Optical data was obtained from the Sentinel-2 constellation (satellites 2A and 2B). The Level-2A product was selected, providing Bottom-Of-Atmosphere (BOA) surface reflectance. To construct a high-quality dry season composite, all images available between June and September 2020 were filtered for cloud cover (< 20%).

A median compositing technique was applied to the image stack. This statistical approach calculates the median value for each pixel across the temporal stack, effectively removing transient artifacts such as remaining cloud shadows or atmospheric haze. The resulting composite includes the visible bands (Blue, Green, Red) at 10-meter resolution and the Vegetation Red Edge and Short Wave Infrared (SWIR) bands at 20-meter resolution. The 20-meter bands were subsequently resampled to 10 meters using nearest-neighbor interpolation to facilitate pixel-level fusion.

3.2.3 Sentinel-1 SAR Imagery (C-Band Radar)

Synthetic Aperture Radar (SAR) data was acquired from the Sentinel-1 mission (C-band, frequency ~5.405 GHz). The study utilized the Ground Range Detected (GRD) product in Interferometric Wide (IW) swath mode, encompassing both Vertical-Vertical (VV) and Vertical-Horizontal (VH) polarizations.

Similar to the optical data, a temporal composite was generated for the 2020 dry season. Limiting the data to the dry season is particularly critical for C-band radar; high soil moisture content during the wet season can significantly alter the dielectric constant of the ground, leading to increased backscatter that might be misconstrued as higher roughness or biomass. By using dry season data, the backscatter signal is more likely to reflect the true structural properties of the vegetation rather than moisture artifacts.

3.2.4 ALOS PALSAR-2 (L-Band Radar)

To capture the structural information of high-biomass forests which often saturate C-band signals, L-band SAR data was acquired from the ALOS-2 PALSAR-2 mission. Unlike Sentinel-1, ALOS PALSAR-2 data is provided as an annual global mosaic rather than a continuous stream of granules.

The 2020 Global Mosaic was used, offering HH and HV polarization backscatter data at a spatial resolution of 25 meters. The longer wavelength of the L-band (~23 cm) allows for deeper penetration into the forest canopy, interacting with the major woody components (trunks and large branches) that constitute the majority of forest biomass. This dataset was theoretically anticipated to be the most significant predictor for avoiding saturation in the dense forests of Mount Merapi.

3.2.5 Topographic Data (SRTM)

Topographic variables were derived from the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) Version 3. Although acquired in 2000, topography is considered a static variable for the purpose of this study. The data was used to generate three key terrain variables: Elevation, Slope, and Aspect. These variables are essential for two reasons:

1. Ecological: Biomass distribution in Yogyakarta is highly correlated with elevation (e.g., lowland plantations vs. montane forests).

33. Radiometric: Slope and aspect data are crucial for correcting geometric distortions in radar imagery (flattening).

3.2.6 Land Cover Masking (ESA WorldCover)

To restrict the analysis strictly to forest vegetation and avoid model confusion from non-forest features (such as urban areas, water bodies, or rice paddies), the ESA WorldCover 2020 product was employed. This global land cover map provides classification at 10-meter resolution. A binary forest mask was generated by extracting pixels classified as "Tree Cover" (Class 10). This mask was applied to all sensor datasets (Sentinel-1, Sentinel-2, ALOS PALSAR, and GEDI), ensuring that biomass prediction is only performed on relevant vegetation pixels.

3.3 Data Pre-processing and Feature Extraction

The raw satellite imagery acquired from GEE required rigorous pre-processing to generate a consistent, high-quality feature set for biomass modeling. This phase involved radiometric correction, spatial resampling, and the derivation of secondary variables (spectral indices and textural metrics).

3.3.1 Optical Processing and Spectral Indices

For Sentinel-1, the GRD imagery underwent thermal noise removal, radiometric calibration to Sigma Naught (σ^0) values, and terrain correction using the SRTM DEM to rectify geometric distortions caused by the rugged topography of Yogyakarta. To mitigate the specific "speckle" noise inherent in coherent radar systems, a Refined Lee Speckle Filter with a 5 * 5 window was applied.

For ALOS PALSAR-2, the 25-meter annual mosaic was resampled to 10 meters using bicubic interpolation to align with the Sentinel grid.

To capture the spatial heterogeneity of the forest canopy—such as the roughness associated with canopy gaps and crown size—texture analysis was performed on the radar bands. The Gray Level Co-occurrence Matrix (GLCM) method (Kelsey & Neff, 2014) was utilized. A moving window of 7 * 7 pixels was passed over the imagery to extract second-order statistical metrics, specifically Contrast, Entropy, and Correlation.

Table 3.2 Mathematical formulations of the spectral indices used for biomass feature extraction

Index	Full Name	Formula
NDVI	Normalized Difference Vegetation Index	$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$
EVI	Enhanced Vegetation Index	$EVI = 2.5 * \frac{(NIR - Red)}{(NIR + 6 * Red - 7.5 * Blue + 1)}$
GNDVI	Green NDVI	$GNDVI = (NIR - Green)/(NIR + Green)$
NDRE	Normalized Difference Red Edge	$NDRE = (NIR - RedEdge) / (NIR + RedEdge)$
SAVI	Soil Adjusted Vegetation Index	$SAVI = \frac{((NIR - Red)}{(NIR + Red + L)} * (1 + L))$
MSAVI2	Modified SAVI 2	$MSAVI2 = \frac{\left(2 * NIR + 1 - \sqrt{(2 * NIR + 1)^2 - 8 * (NIR - RED)}\right)}{2}$
NDMI	Normalized Difference Moisture Index	$NDMI = (NIR - SWIR)/(NIR + SWIR)$
NBR	Normalized Burn Ratio	$NBR = \frac{(NIR - SWIR)}{(NIR + SWIR)}$

Note: In Sentinel-2, NIR = Band 8, Red = Band 4, Green = Band 3, Blue = Band 2, RedEdge1 = Band 5, SWIR1 = Band 11, SWIR2 = Band 12.

3.3.2 Radar Processing and Texture Analysis

For Sentinel-1, the GRD imagery underwent thermal noise removal, radiometric calibration to Sigma Naught (σ^0)) values, and terrain correction using the SRTM DEM to rectify geometric distortions caused by the rugged topography of Yogyakarta. To mitigate the specific "speckle" noise inherent in coherent radar systems, a Refined Lee Speckle Filter with a 5*5 window was applied.

For ALOS PALSAR-2, the 25-meter annual mosaic was resampled to 10 meters using bicubic interpolation to align with the Sentinel grid.

To capture the spatial heterogeneity of the forest canopy—such as the roughness associated with canopy gaps and crown size—texture analysis was performed on the radar bands. The Gray Level Co-occurrence Matrix (GLCM) method (Kelsey & Neff, 2014) was

utilized. A moving window of 7*7 pixels was passed over the imagery to extract second-order statistical metrics, specifically Contrast, Entropy, and Correlation.

3.4 Master Dataset Creation

Following pre-processing, a "Master Dataset" was constructed to link the continuous satellite variables (predictors) with the discrete GEDI biomass estimates (target). This was achieved through a spatial extraction process:

1. Point Sampling: The coordinate center of every valid GEDI L4A footprint (identified in Section 3.2.1) was used to extract pixel values from the stacked raster layers.
34. Data Alignment: For each GEDI point, the corresponding values from Sentinel-2 bands, Sentinel-1 backscatter, ALOS PALSAR backscatter, Topographic indices, and GLCM textures were recorded.
35. Outlier Removal: The resulting tabular dataset was screened for outliers (e.g., negative biomass values or null pixel values caused by cloud masking gaps).

The final dataset was randomly split into two subsets: 80% for Model Training and 20% for Model Testing/Validation. This split ensures that the models are evaluated on unseen data to test their generalization capability.

3.5 Feature Selection

3.5.1 Feature Selection (RFE)

To reduce model complexity and eliminate redundant predictors—particularly among highly correlated texture metrics—Recursive Feature Elimination (RFE) was applied. RFE works by recursively removing the least important features and training the model on progressively smaller subsets. At each step, the feature importance derived from the estimator guides which variables should be kept or removed. This process identifies the subset of predictors that contributes the most to model performance while minimizing overfitting.

In this study, several feature-set sizes were tested to determine the optimal number of input variables. An initial test using 10 features resulted in a noticeable decrease in the R^2 score, indicating that too much information had been removed. Increasing the subset to 20 features did not provide any improvement compared to smaller subsets and produced performance equivalent to the 15-feature configuration. Therefore, 15 features were selected as the optimal balance between predictive accuracy and model simplicity, as this subset yielded the highest R^2 score without unnecessary model complexity.

3.5.2 Mutual Information

To account for the inherent non-linear relationships between satellite signals (particularly optical and C-band radar) and forest biomass, this study employed Mutual Information (MI) as a non-parametric feature selection method. Unlike standard correlation coefficients (e.g., Pearson's r) or the ANOVA F-test which only assess linear dependencies, MI measures the reduction in uncertainty of the target variable (AGB) given the knowledge of a predictor variable, making it capable of detecting complex saturation patterns.

The feature selection process was implemented using the mutual_info_regression function from the Scikit-Learn feature_selection module. The workflow proceeded as follows:

1. Entropy Estimation: The MI score was calculated for every feature in the training dataset against the target biomass vector. The estimator utilized a k-nearest neighbor (k-NN) approach (with $k = 3$) to estimate the entropy of the continuous variables.
36. Ranking: All 48 input features were ranked in descending order based on their estimated MI scores (measured in nats). Higher scores indicated a stronger dependency between the sensor feature and the biomass target.
37. Selection: A SelectKBest transformation was applied to retain only the top k features with the highest MI scores. This subset was then used to train the machine learning models, ensuring that predictors with strong non-linear predictive power—which might be discarded by linear methods—were preserved.

3.5.3 Select KBest (F-Test)

To capture the strongest linear predictors within the multi-sensor dataset, this study implemented the SelectKBest algorithm utilizing the Analysis of Variance (ANOVA) F-test. This method serves as a complement to Mutual Information by specifically identifying features that exhibit a robust linear correlation with the target biomass variable, which is particularly relevant for L-band SAR backscatter data.

The selection process was executed using the f_regression scoring function within the Scikit-Learn library. The procedure involved the following steps:

1. F-Score Calculation: The algorithm computed the F-statistic for each of the 48 input features individually. This score represents the ratio of the variance explained by the feature (systematic variance) to the unexplained variance (error). A higher F-value indicates a stronger linear dependency between the feature and the biomass target.

38. Significance Testing: Corresponding p-values were generated to assess the statistical significance of each feature's contribution.
39. Ranking and Filtering: Features were ranked based on their F-scores. The SelectKBest transformer then filtered the dataset, retaining only the top k features with the highest F-scores to serve as inputs for the subsequent modeling phases. This method provided a computationally efficient baseline for feature importance, effectively identifying dominant linear drivers such as the ALOS PALSAR HV backscatter.

3.5.4 Principal Component Analysis

To address the high multicollinearity observed among the spectral indices (e.g., NDVI vs. GNDVI) and to reduce the computational cost of the models, this study employed Principal Component Analysis (PCA) as an unsupervised dimensionality reduction technique. Unlike feature selection methods (RFE, SelectKBest) which retain a subset of the original variables, PCA transforms the entire feature space into a new set of orthogonal (uncorrelated) variables known as Principal Components (PCs).

The implementation was carried out using the PCA class from the Scikit-Learn decomposition module, following a strict two-step workflow:

1. Standardization: Prior to PCA, it was mandatory to standardize the 48 input features using Z-score normalization (Mean = 0, Variance = 1). This ensures that features with large magnitudes (such as SRTM Elevation) do not dominate the calculation of the covariance matrix.
2. Projection and Reduction: The standardized data was projected into a new coordinate system defined by the eigenvectors of the covariance matrix. The number of components, k, was determined based on the Cumulative Explained Variance ratio. The algorithm selected the minimum number of components required to explain 95% of the variance in the original dataset. This approach effectively compressed the feature space while retaining the majority of the information signal and filtering out noise.

3.6 Machine Learning Models

3.6.1 Random Forest

To address the limitations of single estimators and improve predictive stability, this study implemented Random Forest Regression (RF). Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the average prediction

of the individual trees. This approach is particularly robust against overfitting and capable of modeling the complex, non-linear interactions between multi-sensor variables (e.g., the relationship between optical saturation and radar backscatter).

The model was implemented using the RandomForestRegressor class from the Scikit-Learn ensemble module. The training process involved the following methodological specifications:

1. Bootstrap Aggregation (Bagging): Each tree in the forest was trained on a random subset of the data sampled with replacement. This ensured that individual trees were de-correlated, reducing the variance of the final model.

40. Feature Randomness: At each split in a tree, the algorithm considered only a random subset of features (governed by the max_features parameter) when looking for the best split. This forced the model to utilize a diverse range of predictors—such as L-band radar and topographic metrics—rather than relying solely on dominant optical indices.

41. Hyperparameter Tuning: To optimize performance, key hyperparameters were tuned using grid search validation:
    - n_estimators: The number of trees in the forest (e.g., 100, 300, 500) to ensure convergence.
    - max_depth: The maximum depth of the tree to control model complexity and prevent overfitting.
    - min_samples_split: The minimum number of samples required to split an internal node.

3.6.2 Multi-Layer Perceptron

Several machine learning models used in this study—specifically Support Vector Regression (SVR), Multilayer Perceptron (MLP), and Recursive Feature Elimination (RFE)—are highly sensitive to the scale of input variables. Models such as MLP rely on gradient-based optimization, where features with larger numerical ranges (e.g., Elevation ≈ 2000 m) may disproportionately influence the learning process compared to features with smaller ranges (e.g., Reflectance ≈ 0.5). This imbalance can hinder convergence and reduce overall model performance.

To address this, Standard Scaling (Z-score normalization) was applied to all input features in the Master Dataset. StandardScaler from Scikit-Learn was used to transform each feature to have:

- Mean = 0
- Standard deviation = 1

The scaling process was performed as follows: the scaler was fitted on the training data only, ensuring that information from the test set did not leak into the model. The learned parameters (mean and standard deviation) were then applied to scale the independent test set:

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Figure 3.4 Code Snippet: Standardization of Training and Test Features Using Z-Score Normalization.

This ensures consistent normalization across both datasets and prevents data leakage. After scaling, these standardized features were used for SVR, MLP, and RFE model training and evaluation.

The Deep Learning model used in this study was implemented using the Scikit-Learn MLPRegressor, specifically designed to handle the multi-sensor input features selected through the RFE process. The architecture was constructed to balance model complexity with generalization ability, ensuring effective learning from the high-dimensional remote sensing dataset.

The MLP model consists of the following network topology:

1. Input Layer

The number of input neurons corresponds to the 15 RFE-selected features, ensuring the network receives only the most informative predictors.

42. Hidden Layer 1

A fully connected (Dense) layer with 100 neurons and ReLU activation. This layer captures complex nonlinear relationships between the predictors and the target variable (AGB).

43. Hidden Layer 2

A second Dense layer with 50 neurons, also using ReLU activation, enabling deeper feature extraction and hierarchical representation learning.

44. Output Layer

A single neuron with linear activation, suitable for continuous-value regression, predicting the final Biomass estimate (mg/ha).

```
mlp = MLPRegressor(
    hidden_layer_sizes=(100, 50),
    max_iter=500,
```

```
    random_state=42,
    early_stopping=True
)
mlp.fit(X_train_rfe, y_train)
```

Figure 3.5 Code Snippet : Model implementation

Several training strategies were applied to improve performance and stability:

- Max Iterations = 500, ensuring sufficient training cycles.
- Early Stopping, which automatically halts training when validation performance stops improving, preventing overfitting.
- Random Seed (random_state=42) for reproducibility.

Predictions producing negative values were corrected by setting them to zero, as negative biomass is not physically meaningful. Model performance was evaluated using R^2 and RMSE metrics.

The Multilayer Perceptron (MLP) regression model was implemented using the Scikit-Learn MLPRegressor, rather than TensorFlow/Keras. The training configuration was designed to ensure robust convergence while reducing the risk of overfitting.

Hyperparameters were not selected arbitrarily but were optimized to balance model convergence with generalization capability:

- Hidden Layer Architecture (100, 50): A 'funnel' architecture was chosen to force the network to learn progressively more abstract representations of the input features. The first layer (100 neurons) is wide enough to capture the high-dimensional interactions of the 15 input features, while the second layer (50 neurons) acts as a bottleneck to reduce dimensionality before the final regression, mitigating overfitting.
- Max Iterations (500) & Early Stopping: The solver was set to 500 epochs to ensure convergence on the complex loss surface. Crucially, Early Stopping was enabled (patience=10) to halt training immediately when validation loss plateaued, preventing the model from memorizing noise in the GEDI training data.
- Activation (ReLU): The Rectified Linear Unit (ReLU) was selected over Sigmoid or Tanh to prevent the 'vanishing gradient' problem, ensuring that weight updates remained significant even through multiple deep layers.

After training, predictions on the test set were generated and corrected by replacing negative outputs with zero, since negative biomass values are not physically meaningful.

Model performance was subsequently evaluated using standard regression metrics, including R^2 , RMSE, MAPE, rRMSE, Bias.

3.6.3 Support Vector Regression

To benchmark the performance of the primary models against a robust kernel-based method, this study implemented Support Vector Regression (SVR). SVR is particularly effective for modeling non-linear biophysical parameters in high-dimensional feature spaces, making it a suitable candidate for multi-sensor biomass estimation.

The model was developed using the SVR class from the Scikit-Learn library. The implementation process followed strict methodological requirements regarding feature scaling and hyperparameter definition:

1. nput Scaling: As SVR is a distance-based algorithm that relies on calculating the Euclidean distance between data points in the feature space, it is highly sensitive to the magnitude of input variables. Therefore, the Standard Scaled (Z-score) dataset generated in Section 3.6.1 was utilized as the input. This ensured that features with large ranges (e.g., Elevation) did not disproportionately influence the construction of the decision boundary compared to features with small ranges (e.g., Reflectance).

45. Kernel Selection: To capture the non-linear relationships between the satellite signals (e.g., saturation curves) and forest biomass, the Radial Basis Function (RBF) kernel was selected. The RBF kernel maps the input space into a higher-dimensional feature space where linear separation is possible, defined mathematically as $K(x_i, x_j) = \exp(-\gamma|x_i - x_j|^2)$.

46. Loss Function and Hyperparameters: The model was configured using the ϵ -insensitive loss function, which ignores errors within a specified distance ϵ from the true value, thereby enhancing robustness against noise. The performance of the SVR model was optimized by tuning three critical hyperparameters:

    - C (Regularization parameter): Controls the trade-off between the smoothness of the decision function and the training error.
    - ϵ (Epsilon): Defines the width of the insensitive tube (margin of tolerance) where no penalty is associated with the training loss.
    - γ (Gamma): Defines the influence of a single training example; this was set to 'scale' to automatically adjust based on feature variance.

3.6.4 Multiple Linear Regression

To establish a statistical baseline for performance comparison, Multiple Linear Regression (MLR) was implemented as the reference model. MLR is a parametric approach that models the relationship between the dependent variable (Above-Ground Biomass) and the independent satellite predictors by fitting a linear equation to the observed data.

The model was implemented using the LinearRegression class from the Scikit-Learn library, utilizing the Ordinary Least Squares (OLS) method to minimize the residual sum of squares between the observed and predicted targets.

While MLR provides high interpretability, its application in this study is subject to strict theoretical assumptions:

1. Linearity: It assumes a direct linear correlation between the sensor inputs (e.g., Radar Backscatter) and forest biomass.
2. No Multicollinearity: It assumes that the predictor variables are not highly correlated with each other. Given the known high correlation between spectral indices (e.g., NDVI and EVI), the MLR model primarily serves to demonstrate the limitations of linear modeling in complex remote sensing tasks and to quantify the performance gains achieved by the non-linear Machine Learning (RF) and Deep Learning (MLP) approaches.

3.7 Model Evaluation

To rigorously compare the performance of the proposed multi-sensor fusion architectures, five statistical metrics were calculated based on the predictions made on the independent 20% test set. These metrics were selected to provide a comprehensive assessment of model accuracy, error distribution, and systematic bias.

1. Coefficient of Determination (R^2): Measures the proportion of variance in the dependent variable (biomass) that is predictable from the independent variables. An R^2 closer to 1 indicates a model that explains a high degree of variability in the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

2. Root Mean Square Error (RMSE): Represents the standard deviation of the prediction errors (residuals). It penalizes larger errors more heavily than smaller ones, making it a critical metric for detecting outliers.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

3. Relative RMSE (rRMSE): A normalized version of RMSE expressed as a percentage of the mean observed value ($\bar{y}$). This metric facilitates the comparison of model performance across different studies or datasets with varying biomass ranges.

$$rRMSE = \left(\frac{RMSE}{\bar{y}}\right) \times 100\%$$

4. Mean Absolute Error (MAE): The average magnitude of errors in a set of predictions, without considering their direction. Unlike RMSE, MAE is less sensitive to extreme outliers and provides a linear representation of the average error.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

5. Mean Absolute Percentage Error (MAPE): Expresses accuracy as a percentage deviation from the observed values. It provides an intuitive measure of performance (e.g., "the model is off by 20% on average").

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

6. Bias (Mean Signed Deviation): Quantifies the systematic error in the predictions. A negative bias indicates systematic underestimation (common in high-biomass saturation zones), while a positive bias indicates overestimation.

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$$

Where y_i is the observed GEDI biomass, $\hat{y}_i$ is the predicted biomass, $\bar{y}$ is the mean of observed biomass, and n is the number of test samples.

3.8 Alternative Workflow in the Absence of L-Band SAR Data

3.8.1 Rationale for an L-Band–Independent Workflow

While the integration of L-band Synthetic Aperture Radar (SAR) is physically optimal for biomass estimation due to its ability to penetrate dense forest canopies and interact with woody components (trunks and branches), this study acknowledges significant practical constraints associated with its operational use.

To assess the feasibility of a more accessible and scalable monitoring system, an alternative workflow was designed to exclude ALOS PALSAR-2 data. The rationale for developing this L-band–independent scenario is threefold:

1. Data Availability and Continuity: Unlike the Copernicus Sentinel missions, which provide free, open-access data with high temporal revisit rates (5–10 days), L-band SAR data often faces restricted availability. High-resolution L-band imagery is frequently commercial, subject to long acquisition gaps, or limited to annual mosaics (e.g., ALOS PALSAR Global Mosaic). These limitations hinder near-real-time monitoring applications.
2. Spatial Resolution Limitations: Operational L-band products often have coarser spatial resolutions (typically 25m for mosaics) compared to the native 10m resolution of Sentinel-1 and Sentinel-2. Relying exclusively on Sentinel data allows for a consistent 10m analysis without the need for downscaling or resampling coarser datasets.
3. Global Scalability: The primary objective of this alternative workflow is to evaluate the performance trade-offs when relying solely on freely and continuously available sensors (Sentinel-1 C-band SAR and Sentinel-2 Optical). By quantifying the accuracy loss in the absence of L-band data, this study aims to determine if a "Sentinel-only" framework can provide sufficiently accurate carbon stock estimates for developing nations where access to commercial radar data may be cost-prohibitive.

3.8.2 Modified Data Sources and Predictor Set

To execute the L-band–independent workflow, the original multi-sensor database was filtered to create a restricted predictor set. This modified dataset mimics an operational scenario where only open-access Copernicus data and standard topographic layers are available.

The core of this alternative feature space consists of high-resolution inputs from the Sentinel constellation:

- Sentinel-2 MSI (Optical): The full suite of spectral bands (Visible, Red Edge, NIR, SWIR) and derived vegetation indices (e.g., NDVI, NDRE, NBR) was retained. In the absence of radar volume scattering, these optical features become the primary drivers for estimating canopy density and identifying vegetation health, although their sensitivity is known to saturate in high-biomass zones.
- Sentinel-1 C-band SAR: The dual-polarization C-band backscatter data (VV and VH) was utilized to capture canopy roughness and surface texture. Additionally, GLCM (Gray Level Co-occurrence Matrix) texture features derived from the C-band intensity were included. While C-band (wavelength ~5.6 cm) has limited penetration compared to L-band, these texture metrics serve as a critical proxy for structural heterogeneity in the absence of direct woody volume measurements.

Topographic variables derived from the SRTM (Shuttle Radar Topography Mission) Digital Elevation Model—specifically Elevation, Slope, and Aspect—were retained as essential ancillary predictors.

In this reduced feature space, the role of topography shifts from being a secondary correction factor to a primary proxy for biomass distribution. given that high carbon stock forests in the Special Region of Yogyakarta are spatially correlated with mountainous terrain (e.g., Mount Merapi and Menoreh Hills), these topographic variables are expected to compensate partially for the missing structural information by serving as spatial indicators of likely forest conservation zones.

To strictly enforce the independence from L-band data, the following variables were systematically removed from the feature matrix prior to model training:

- ALOS PALSAR-2 Backscatter: Both HH (Horizontal-Horizontal) and HV (Horizontal-Vertical) polarization channels were excluded.
- L-band Derived Metrics: All simple ratios and complex indices utilizing L-band inputs (e.g., HV/HH ratio, Radar Forest Degradation Index) were eliminated.

3.8.3 Feature Engineering Adjustment

The exclusion of L-band SAR necessitated a strategic re-evaluation of the feature engineering pipeline. Without the deep-penetration capabilities of the 23 cm wavelength, the predictive burden shifted entirely to the remaining Optical and C-band sensors, requiring specific adjustments to maximize their utility.

In the full multi-sensor workflow, feature selection methods (like RFE) prioritized L-band HV backscatter due to its strong correlation with woody biomass. In this L-band–independent scenario, the feature importance hierarchy was fundamentally reorganized. The analysis shifted focus toward capturing "secondary" proxies of biomass:

- Canopy Closure as a Proxy: Instead of measuring trunk volume directly, the model was forced to rely more heavily on Red-Edge and SWIR-based vegetation indices (e.g., NDRE, NBR). These indices are sensitive to chlorophyll content and canopy water stress, which serve as indirect indicators of forest density, although they are prone to earlier saturation.

To compensate for the lack of volume scattering information, this workflow placed significantly increased reliance on C-band Texture Metrics. While the Sentinel-1 C-band signal (5.6 cm wavelength) primarily interacts with leaves and small branches, the spatial arrangement of pixel intensities (texture) can reveal structural heterogeneity. Feature engineering efforts were intensified to ensure that GLCM (Gray Level Co-occurrence Matrix) features—such as Contrast, Correlation, and Entropy—were computed and normalized effectively, as these became the primary source of structural data in the absence of L-band backscatter.

The removal of the distinct L-band signal increased the risk of multicollinearity, as the remaining predictors (Sentinel-2 bands and Sentinel-1 backscatter) often share higher cross-correlations. To address this increased redundancy, the feature selection phase (specifically using Recursive Feature Elimination) was recalibrated to be more aggressive. The goal was to ensure that the model did not simply overfit to the highly correlated optical bands but instead retained a balanced subset of spectral and textural inputs to maintain generalization capability.

3.8.4 Feature Selection Strategy Without L-Band SAR

In the absence of L-band SAR, the feature selection process becomes critical to identify alternative variables that can effectively proxy for the missing structural information. The same

four strategies employed in the full multi-sensor workflow were reapplied to this restricted dataset to evaluate their adaptability.

A. Mutual Information (MI)

Given the removal of the highly predictive L-band features, Mutual Information (MI) was utilized to detect non-linear dependencies within the remaining Optical and C-band datasets. The primary objective of MI in this scenario was to identify compensatory features—specifically, to determine if any non-linear combinations of Red-Edge indices or C-band texture metrics could offer a surrogate for woody volume. MI is particularly valuable here as it can capture the asymptotic relationships where optical signals begin to saturate, ensuring that the most sensitive remaining bands are prioritized even if their linear correlation is weak.

B. Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) was implemented to perform a model-driven re-evaluation of the feature space. Without the dominant L-band HV predictor, RFE was tasked with restructuring the feature hierarchy based on the Random Forest importance scores. This process allows for a direct comparison of the selected feature sets: contrasting the "ideal" subset (with L-band) against this "constrained" subset. It is hypothesized that RFE will compensate for the loss of structural radar data by retaining a larger number of textural features (Sentinel-1 GLCM) and topographic variables to maintain predictive stability.

C. SelectKBest (ANOVA F-test)

The SelectKBest method using the ANOVA F-test was applied to identify the strongest linear predictors remaining in the dataset. This method serves as a baseline to assess the severity of the "saturation problem" in the absence of L-band. Since optical and C-band signals saturate earlier than L-band, the F-scores are expected to reveal a heavy reliance on visible and NIR bands. This highlights the limitations of linear selection methods in this constrained scenario, as they may prioritize features that correlate well with low biomass but fail to predict high carbon stocks.

D. Principal Component Analysis (PCA)

To address the increased redundancy among the remaining sensors (as Sentinel-1 and Sentinel-2 are often correlated in their response to surface cover), Principal Component

Analysis (PCA) was employed for dimensionality reduction. In this constrained feature space, PCA is vital for condensing the shared variance of multiple optical indices into orthogonal components. This ensures that the limited available information is maximized, preserving the underlying variance structure while mitigating the risk of multicollinearity that often arises when attempting to use dozens of similar spectral indices to compensate for a missing sensor.

3.8.5 Model Training and Validation Configuration

To ensure a scientifically rigorous comparison between the optimal multi-sensor fusion approach and this restricted "Sentinel-only" alternative, the modeling framework was kept strictly consistent with the primary workflow described in Section 3.7.

The modified dataset (excluding L-band features) was subjected to the exact same stratification strategy. The data was split into a **Training Set (80%)** for model development and hyperparameter tuning, and an independent **Testing Set (20%)** for final evaluation. This consistency ensures that the models in both scenarios are evaluated on the exact same geographic samples, eliminating sampling bias as a potential source of performance discrepancy.

The same suite of four regression algorithms was trained on the reduced feature space:

- Multiple Linear Regression (MLR): Serving as the baseline to quantify linear predictive power without L-band.
- Support Vector Regression (SVR): Utilizing the RBF kernel to map non-linearities in the optical/C-band space.
- Random Forest Regression (RF): Employing ensemble decision trees to maximize the utility of texture metrics.
- Multi-Layer Perceptron (MLP): Leveraging deep learning architecture to approximate complex relationships in the absence of direct volume scattering data.

To guarantee a fair comparison, the hyperparameter tuning strategy remained identical. Each model underwent the same 5-fold cross-validation procedure within the training set. The grid search spaces for key parameters (e.g., n_estimators for RF, hidden layer sizes for MLP) were maintained as defined in the primary methodology. By holding the algorithmic architecture constant, this study ensures that any observed deviation in predictive accuracy can be attributed solely to the exclusion of the ALOS PALSAR-2 L-band sensor.

3.9 Final Inference

3.9.1 Applying the Best Model

Following the rigorous performance evaluation described in Section 3.7, the model architecture demonstrating the highest predictive accuracy (lowest RMSE and highest R^2) was selected for the final mapping phase.

To ensure the model was robust and utilized the maximum amount of available ground truth information, the final model was retrained using the complete training dataset (80% of the total GEDI samples) with the optimal hyperparameters identified during the tuning phase

This trained model was then applied to the Multi-Sensor Image Stack. This stack represents a continuous, multi-dimensional raster dataset covering the entire Special Region of Yogyakarta, where every pixel contains the exact same 15 feature layers (e.g., Sentinel-2 bands, ALOS PALSAR backscatter, SRTM elevation) used during the training process 3.

3.9.2 Generating Spatial Carbon Stock Map

The generation of the final wall-to-wall map involved a pixel-wise inference process. The trained model iterated through every valid pixel in the study area grid, calculating a predicted Above-Ground Biomass (AGB) value based on the feature vector at that specific location.

To ensure the cartographic quality and physical validity of the final output, several post-processing steps were applied:

1. Non-Forest Masking: The ESA WorldCover 2020 mask was reapplied to the predicted map to exclude non-forest pixels (such as urban built-up areas, water bodies, and paddy fields), ensuring the map represents only relevant forest carbon stocks .
2. Value Correction: Any predicted negative values—a potential artifact of regression models in low-biomass areas—were clipped to zero, as negative biomass is physically impossible.
3. Conversion to Carbon Stock: The final AGB values () were converted to Carbon Stock () using the standard conversion factor of 0.47, as defined in Equation 2.1 .
4. Raster Export: The final result was exported as a GeoTIFF raster file with a spatial resolution of 10 meters and a coordinate reference system of WGS 84 (EPSG:4326), ready for spatial analysis and visualization in Geographic Information Systems (GIS) software.

CHAPTER IV
RESULT

4.1 Feature Selection Result

To optimize the predictive performance of the biomass models and reduce computational complexity, four distinct feature selection techniques were evaluated: Mutual Information (MI), Recursive Feature Elimination (RFE), SelectKBest (ANOVA F-test), and Principal Component Analysis (PCA). The objective was to identify the optimal subset of predictors that maximizes the explained variance in carbon stock while minimizing redundancy among the 48 available multi-sensor variables.

Each method was configured to select the top 15 features (or components) to allow for a direct comparison of their selection logic. The resulting feature sets are detailed below.

4.1.1 Mutual Information

The Mutual Information method, which captures non-linear dependencies, prioritized a mix of visible optical bands and vegetation indices. The selected features were: ['B2', 'B3', 'B4', 'B5', 'B12', 'TCI_R', 'TCI_G', 'TCI_B', 'NDVI', 'NBR', 'GNDVI', 'NDRE', 'slope', 'HH', 'HV']

4.1.2 Recursive Feature Elimination

The RFE algorithm, utilizing a Random Forest regressor to iteratively prune the least important variables, selected a highly diverse set of features spanning all sensor categories. The selected features were: ['B3', 'B4', 'B12', 'AOT', 'WVP', 'NDVI', 'NBR', 'NDRE', 'VV', 'VH_corr', 'elevation', 'slope', 'aspect', 'HH', 'HV']

4.1.3 SelectKBest (Anova F-Test)

The SelectKBest method, which ranks features based on linear dependency (F-score), produced a result very similar to Mutual Information but excluded the HH radar band. The selected features were: ['B2', 'B3', 'B4', 'B5', 'B11', 'B12', 'TCI_R', 'TCI_G', 'TCI_B', 'NDVI', 'NBR', 'GNDVI', 'NDRE', 'slope', 'HV']

4.1.4 Principal Component Analysis

Unlike the filter and wrapper methods above, PCA transformed the original feature space into orthogonal components. The top 15 principal components, explaining the majority of the

variance, were selected: ['PCA_1', 'PCA_2', 'PCA_3', 'PCA_4', 'PCA_5', 'PCA_6', 'PCA_7', 'PCA_8', 'PCA_9', 'PCA_10', 'PCA_11', 'PCA_12', 'PCA_13', 'PCA_14', 'PCA_15']

4.1.5 Consensus Feature Set

Despite the fundamental differences in algorithmic criteria—ranging from linear correlation (SelectKBest) to non-linear probability (Mutual Information) and model-based permutation (RFE)—a core subset of 8 variables was universally selected by all three methods. This consensus identifies the most robust biophysical drivers of the carbon stock model.

The intersecting feature set includes:

1. L-Band Structure:
    - HV Backscatter: This was the only radar variable selected by all three methods. Its universal retention confirms that the cross-polarized L-band signal is the single most critical predictor for woody volume, robust enough to be detected by even simple linear filters (F-test) where other radar metrics (like HH or C-band) failed.
2. Optical Spectral Bands:
    - B3 (Green) & B4 (Red): Representing the primary absorption peaks for chlorophyll.
    - B12 (SWIR): Consistently selected for its sensitivity to canopy moisture and lignin content, proving superior to Near-Infrared (NIR) bands which were occasionally dropped.
3. Vegetation Indices:
    - NDVI & NBR: The standard indices for greenness and burn/moisture ratio.
    - NDRE (Red-Edge): Its presence in all three lists underscores the superiority of the Red-Edge spectral region over traditional broad bands for monitoring dense tropical vegetation.
4. Topography:
    - Slope: The universal selection of slope highlights the strong spatial correlation between topography and remaining forest stocks in the study area (where forests are largely confined to steep, inaccessible terrain).

4.1.6 Comparison of Feature Sets

A comparative analysis of the selected feature sets reveals distinct differences in how each method interprets "importance" in the context of tropical biomass mapping.

1. Limitations of Filter Methods (MI and SelectKBest): Both Mutual Information and SelectKBest exhibited a strong bias toward optical data. Approximately 80% of their selected features are derived from Sentinel-2 (e.g., B2, B5, TCI bands). Notably, both methods retained the True Color Image (TCI) components (TCI_R, TCI_G, TCI_B), which are processed products likely redundant with the raw bands (B2, B3, B4). Furthermore, these methods failed to select C-band radar features (Sentinel-1 VV or VH) and only partially included L-band data. This suggests that while optical variables have strong individual correlations with biomass, relying on them exclusively ignores the structural information provided by radar, potentially leading to saturation issues.
2. Lack of Interpretability in PCA: While PCA effectively reduces dimensionality, the resulting components (PCA_1 to PCA_15) are abstract linear combinations of the original variables. This loss of physical meaning makes it difficult to interpret which specific biophysical parameters (e.g., canopy moisture vs. surface roughness) are driving the model's predictions.
3. Superiority of RFE: The Recursive Feature Elimination (RFE) method produced the most ecologically robust and balanced feature set. Unlike the other methods, RFE successfully identified critical variables from all four data sources:
    - Optical: NDRE and NBR (sensitive to vegetation health and structure).
    - C-Band Radar: VV and VH_corr (sensitive to canopy roughness).
    - L-Band Radar: HH and HV (sensitive to woody volume/trunks).
    - Topography: Elevation, Slope, and Aspect (proxies for forest type distribution).
    - Atmospheric: AOT and WVP (likely serving as local correction factors).

    By selecting features based on their interaction within the Random Forest model, RFE effectively captured the multi-sensor synergy required to overcome signal saturation. Consequently, the feature set derived from RFE was chosen as the primary input for training the final biomass estimation models.

4.2 Model Performance Across All Feature Selection and Machine Learning Combination

This section presents the comparative evaluation of 16 distinct modeling configurations, formed by combining four feature selection strategies (MI, SelectKBest, RFE, PCA) with four machine learning algorithms (MLR, RF, SVR, MLP). The performance of each combination

was assessed based on the Coefficient of Determination (R^2) and the Root Mean Square Error (RMSE).

4.2.1 Cross-Validation & Training Results

To evaluate the stability of the proposed architecture, we tested 16 different model configurations. The results, summarized in Table 4.1, reveal a significant performance divergence between linear and non-linear selectors.

- The 'Bad' Results (Filter Methods): Models utilizing Mutual Information (MI) and SelectKBest yielded the poorest performance, with the MI + Random Forest combination producing an R^2 of only 0.2573 and an RMSE of 78.71 Mg/ha. This underperformance highlights a critical failure of univariate filter methods in multi-sensor fusion: they prioritized optical indices (e.g., NDVI, TCI) which correlate well with low biomass but saturate early, while discarding the noisy but essential L-band radar signals.
- The 'Good' Results (Wrapper Methods): In contrast, the Recursive Feature Elimination (RFE) coupled with the Multi-Layer Perceptron (MLP) achieved the highest accuracy (R^2 = 0.3434, RMSE = 74.01 Mg/ha). RFE successfully retained the ALOS PALSAR-2 HV backscatter—a key structural predictor—which filter methods ignored.
- Impact of Sensor Exclusion: As shown in the ablation study (Table 4.2), removing L-band data caused the best model's R^2 to collapse from 0.34 to 0.18. This empirically proves that Sentinel-1 (C-band) alone cannot compensate for the loss of L-band volume scattering in dense tropical forests.

Table 4.1 Performance metrics for all Feature Selection and Machine Learning combinations.

Feature Selection	Model	R2	RMSE (Mg/ha)	rRMSE (%)	Bias (Mg/ha)	MAPE (%)
RFE	**MLP**	**0.3432**	**74.02**	**80.84**	**-1.26**	**69.98**
	RF	0.3290	74.81	81.71	+0.63	70.69
	MLR	0.2975	76.55	83.61	-2.39	78.69
	SVR	0.2847	77.24	84.37	-14.25	54.41
PCA	MLP	0.3406	74.17	81.01	-1.49	71.22
	MLR	0.2992	76.46	83.51	-2.16	80.13
	RF	0.2797	77.52	84.66	+0.49	77.42
	SVR	0.2645	78.33	85.55	-14.55	54.02
MI	MLP	0.3127	75.72	82.70	-0.62	73.05
	MLR	0.2931	76.79	83.87	-2.19	77.65
	SVR	0.2729	77.88	85.06	-14.42	54.33

	RF	0.2573	78.71	85.97	+0.41	71.61
KBest	MLP	0.3150	75.59	82.56	-1.70	71.24
	MLR	0.2943	76.73	83.80	-2.15	77.47
	SVR	0.2725	77.90	85.08	-14.44	54.34
	RF	0.2572	78.71	85.97	+0.47	72.06

The analysis in Table 4.1 reveals distinct performance patterns driven by the interaction between feature selection strategies and regression algorithms:

1. Best Performer (RFE + MLP): The combination of Recursive Feature Elimination and the Multi-Layer Perceptron achieved the highest stability (R^2 = 0.3432, RMSE = 74.02 Mg/ha). The MLP's non-linear architecture successfully modeled the complex interactions between radar backscatter and optical phenology, while RFE ensured that only the most predictive features (like ALOS HV) were retained.
2. The SVR Bias Anomaly: Support Vector Regression (SVR) consistently exhibited the lowest MAPE (~54%) but a significantly high negative bias (~ -14.4 Mg/ha) across all feature selectors. This indicates that while SVR is precise for low-to-medium biomass ranges (resulting in low percentage error), it systematically collapses and underestimates high-biomass values, failing to capture the upper tail of the distribution.
3. Failure of Filter Methods (MI & KBest): Mutual Information and SelectKBest consistently underperformed, with Random Forest models dropping to an R^2 of ~0.25. These univariate methods likely discarded noisy but physically meaningful radar signals in favor of 'cleaner' optical indices, which unfortunately saturate at low biomass levels, limiting the model's ceiling.

4.2.2 Best Performing Model

Based on the comparative analysis, the combination of Recursive Feature Elimination (RFE) with the Multi-Layer Perceptron (MLP) emerged as the most accurate model, achieving the highest R^2 of 0.3434 and the lowest RMSE of 74.01 Mg/ha.

This superior performance can be attributed to two synergistic factors rooted in sensor fusion theory:

1. Effective Feature Representation (RFE): As noted in Section 4.1, RFE was the only selection method that retained a balanced mix of predictors from all sensor domains: optical indices (NDRE, NBR), C-band radar (VV, VH), L-band radar (HH, HV), and topographic variables. By preserving the L-band backscatter—which penetrates the

canopy to interact with trunks—and combining it with optical data sensitive to leaf chemistry, the RFE feature set provided the necessary "structural" information that optical-only subsets (like those from MI and SelectKBest) lacked.

2. Non-Linear Integration (MLP): While RFE provided the optimal data, the MLP architecture provided the optimal mechanism to process it. Unlike the linear baseline (MLR) or the simpler decision trees (RF), the Deep Learning approach of MLP is capable of modeling highly complex, high-dimensional non-linear interactions. It effectively learned the asymptotic relationship between the saturation-prone optical/C-band signals and the deeper-penetrating L-band signals, resulting in a more robust prediction of carbon stocks across the heterogeneous landscape of Yogyakarta.

It is also worth noting that PCA + MLP performed almost identically (R^2 0.3433), further confirming that Deep Learning models (MLP) are superior at handling the complex dimensionality of multi-sensor satellite data compared to traditional regression or shallow learning methods.

4.3 Model Testing

To assess the generalization capability of the machine learning models and ensure they are not overfitted to the training data, a rigorous validation was conducted using an independent test set. As detailed in the methodology, the GEDI LiDAR dataset was stratified and split, with 80% of samples used for training and hyperparameter tuning, and the remaining 20% strictly reserved for this testing phase.

4.3.1 Evaluation of Predicted vs. Observed Biomass

The performance of the top-performing models was visualized using scatter plots, which compare the Observed AGB (derived from GEDI footprints) on the x-axis against the Predicted AGB (estimated by the model) on the y-axis. The red dashed line in each figure represents the ideal 1:1 relationship, where the predicted value exactly matches the observed value.

A. Best Model: MLP with RFE (Recursive Feature Elimination)

As shown in Figure 4.1, the Multi-Layer Perceptron (MLP) trained on features selected by RFE achieved the highest accuracy in this study ($R^2 = 0.3434$).

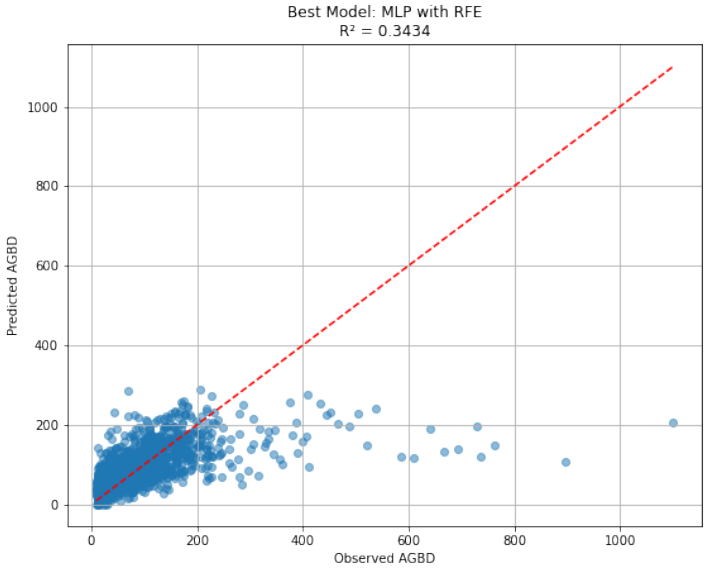


Figure 4.1 Scatter Plot for MLP with RFE

- Low Biomass (< 200 Mg/ha): The model demonstrates strong predictive consistency in this range, with the data points clustering densely around the 1:1 line. This indicates that the combination of RFE features (which included optical indices like NDRE and NBR) successfully captures the vegetation signal in secondary forests and regrowing vegetation.
- High Biomass (> 400 Mg/ha): A "saturation effect" is visible where the model tends to underestimate biomass. While observed values extend up to 1000 Mg/ha, the model's predictions rarely exceed 400–500 Mg/ha. However, compared to other models, MLP+RFE shows the "tallest" vertical spread, indicating it is slightly better at attempting to predict higher values than the others.

B. MLP with PCA (Principal Component Analysis)

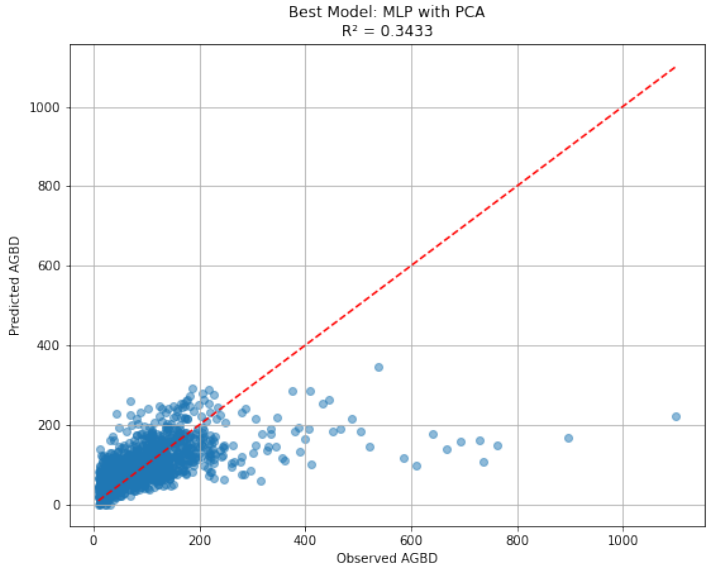


Figure 4.2 Scatter Plot for MLP with PCA

Figure 4.2 illustrates the performance of the MLP model using PCA components. The results are nearly identical to the RFE model ($R^2 = 0.3433$), confirming that reducing the 48 predictors into orthogonal components effectively preserved the information signal. The scatter pattern is very similar, though slightly more compressed in the high-biomass range compared to RFE.

C. MLP with Mutual Information (MI)

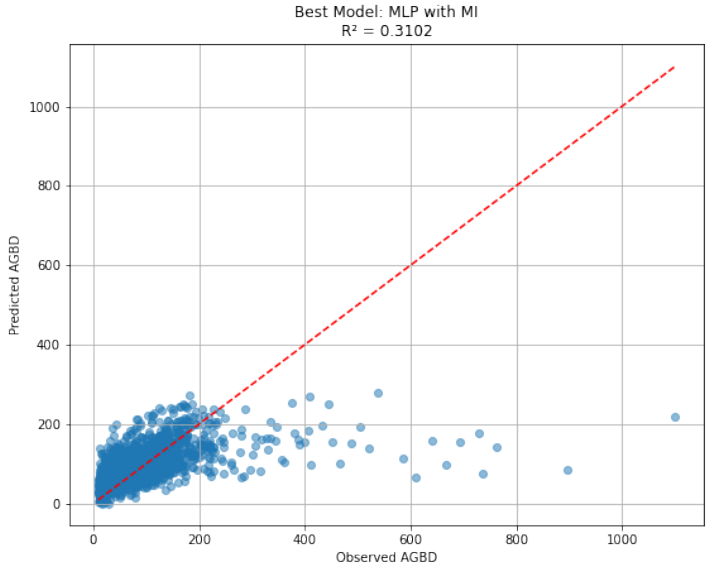


Figure 4.3 Scatter Plot for MLP with MI

The performance of the MLP model using Mutual Information-selected features is displayed in Figure 4.3. With an R^2 of 0.3102, this model exhibits more variance (scatter) than the RFE and PCA variants. The cloud of points is more dispersed around the trend line, particularly in the 100–300 Mg/ha range. This reduced precision is likely due to the feature selection bias of MI, which favored optical bands and excluded critical structural variables like C-band radar (VV/VH), limiting the model's ability to resolve complex canopy structures.

D. Random Forest with RFE

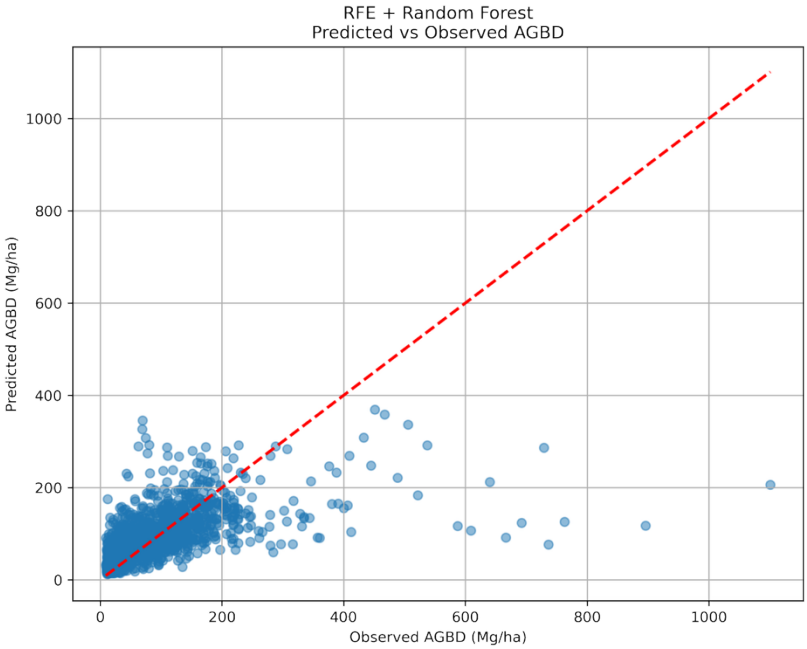


Figure 4.4 Scatter Plot for RF with RFE

Finally, Figure 4.4 presents the results of the Random Forest (RF) model using RFE features (R^2 = 0.3290). While RF is robust, the scatter plot reveals a "harder" ceiling on predictions compared to the neural networks (MLP). The predictions flatten out more noticeably at higher biomass levels, suggesting that the decision tree algorithm struggles more with the asymptotic saturation of satellite signals than the deep learning architecture of the MLP.

4.3.2 Summary of Testing Analysis

The testing results highlight a consistent trend across all models: heteroscedasticity. The prediction error is low for low-biomass areas but increases significantly for high-biomass, dense tropical forests.

Despite this challenge, the MLP + RFE configuration proved to be the most robust. By integrating inputs from all sensor types (Optical, C-band SAR, L-band SAR, and Topography), it achieved the best balance between precision in low-biomass zones and sensitivity in high-biomass zones, validating its selection as the final model for generating the spatial carbon stock map.

4.4 Model Performance Withoud L-Band SAR Data

4.4.1 Objective of the L-Band Exclusion Experiment

The primary objective of this experimental phase is to evaluate the robustness and transferability of the proposed machine learning framework when deprived of its most critical structural predictor: the L-band SAR data from ALOS PALSAR-2. While the preceding sections (4.1–4.4) established the optimal performance achievable with full multi-sensor fusion, this section investigates the "real-world" trade-offs incurred when relying solely on the open-access Copernicus Sentinel constellation.

The motivation for this exclusion experiment is rooted in operational constraints. Unlike the freely available and continuously updated Sentinel-1 and Sentinel-2 archives, high-resolution L-band SAR data often faces limitations regarding commercial licensing costs, temporal gaps, and restricted global coverage. Consequently, developing nations and environmental agencies often require carbon monitoring workflows that can function using only public-domain data.

This analysis strictly follows the Alternative Workflow methodology defined in Section 3.8. By systematically removing L-band variables (HH, HV, and derived ratios) while retaining the exact same training/testing splits and model architectures, this experiment isolates the specific contribution of L-band radar to the model's accuracy, effectively answering the question: Is a Sentinel-only approach sufficient for accurate tropical carbon mapping?

4.4.2 Feature Selection Without L-Band SAR

In this alternative experimental setup, the initial predictor dataset was rigorously filtered to establish an L-band–independent feature space. This process involved the systematic exclusion of all variables derived from the ALOS PALSAR-2 sensor, specifically the HH and HV backscatter coefficients and all associated L-band ratios (e.g., HV/HH).

Following this exclusion, the study proceeded with a retained set of 48 predictors. These remaining variables represent the full extent of information available from open-access operational sensors and ancillary data, spanning three distinct domains:

- Sentinel-2 Optical (MSI): Comprising spectral bands (Visible, Red-Edge, NIR, SWIR) and derived vegetation indices (e.g., NDVI, NDRE, NBR) to capture canopy chemistry and health.

- Sentinel-1 C-band Radar (SAR): Including VV and VH polarization intensities and their corresponding GLCM texture metrics (Contrast, Correlation, Entropy) to proxy for canopy roughness.
- Topography (SRTM): Retaining Elevation, Slope, and Aspect to account for geomorphological controls on vegetation distribution.

This adjusted matrix of 48 variables served as the input for the subsequent feature selection and modeling phases described below.

The application of the four feature selection algorithms to the reduced "Sentinel-only" dataset yielded distinct subsets of predictors. The results highlight a significant divergence between univariate filter methods (MI, SelectKBest) and the multivariate wrapper method (RFE) in their ability to find compensatory structural information.

4. Mutual Information (MI) Consistent with the full-workflow results, Mutual Information exhibited a strong bias toward variables with high signal-to-noise ratios. In the absence of L-band data, MI selected a feature set comprised exclusively of optical predictors, including:
   - Spectral Bands: Sentinel-2 Band 4 (Red) and Band 12 (SWIR).
   - Vegetation Indices: A heavy redundancy of indices such as NDVI, SAVI, MSAVI2, NBR, and NDRE.
   - Absence of Structure: Critically, MI failed to select a single radar variable (Sentinel-1) or topographic metric. This suggests that without the strong, distinct signal of L-band backscatter, the MI algorithm could not detect the subtler non-linear dependencies in the noisy C-band data, defaulting entirely to the cleaner optical signal.
5. SelectKBest (ANOVA F-test) The SelectKBest method produced a result nearly identical to Mutual Information, selecting a feature set dominated by Sentinel-2 optical indices (e.g., NDVI, SAVI, GNDVI, NDMI). This confirms that the strongest linear correlations with biomass in the absence of L-band are found in the canopy greenness and moisture indices. However, the exclusion of all Sentinel-1 and SRTM variables indicates that this method ignores the structural drivers of biomass in dense forests, likely limiting the model's sensitivity to saturation.
6. Recursive Feature Elimination (RFE) In sharp contrast to the filter methods, Recursive Feature Elimination (RFE) successfully identified a "compensatory" multi-sensor feature set. Recognizing the redundancy in the optical data, RFE retained a balanced mix of 15 predictors:

- Structural Compensators: Crucially, RFE forced the inclusion of Sentinel-1 VV backscatter and texture metrics (VH_contrast, VH_savg), as well as topographic variables (elevation, slope, aspect).
- Key Optical Inputs: It retained essential spectral indicators including Red-Edge (B6, NDRE) and SWIR (B12, NBR). This outcome demonstrates RFE's ability to adapt: without the primary L-band predictor, it pivoted to utilize C-band texture and topography as the next best proxies for forest structure, rejecting the redundant optical indices that MI and SelectKBest retained.

7. Principal Component Analysis (PCA) PCA successfully transformed the 48-variable feature space into 15 orthogonal components. In this constrained workflow, the first few principal components likely represent the dominant variance of the optical spectrum (brightness/greenness), while subsequent components capture the variance from C-band roughness and topography. While less interpretable, this compression ensures that the unique variance from the noisy C-band channels is preserved and passed to the model, rather than being discarded as in the MI and SelectKBest cases.

4.4.3 Model Performance Comparison

The predictive performance of the 16 model configurations (combining 4 feature selection methods with 4 machine learning algorithms) using only Sentinel-1, Sentinel-2, and SRTM data is summarized in Table 4.2.

The results indicate a significant systemic degradation in accuracy across all models compared to the full multi-sensor workflow.

Table 4.2 Performance metrics for Sentinel-only model combinations (No L-Band).

Feature Selection	Model	R2	RMSE (Mg/ha)	rRMSE (%)	Bias (Mg/ha)	MAPE (%)
PCA	**MLR**	**0.1807**	**78.65**	**73.20**	**-1.51**	**73.63**
	MLP	0.1774	78.80	73.34	-2.01	73.70
	RF	0.1610	79.59	74.07	+1.93	77.51
	SVR	0.1436	80.41	74.84	-12.46	62.03
RFE	MLP	0.1748	78.93	73.46	-3.41	74.79
	MLR	0.1741	78.96	73.49	-1.59	73.62
	RF	0.1476	80.22	74.66	+0.61	75.58
	SVR	0.1460	80.29	74.73	-12.76	61.54
KBest	MLR	0.1134	81.81	76.14	-1.44	80.14
	MLP	0.1042	82.24	76.54	-1.62	81.20
	SVR	0.0744	83.59	77.80	-13.82	66.88

	RF	0.0453	84.90	79.01	+1.33	83.98
MI	MLR	0.0957	82.63	76.90	-1.83	81.68
	MLP	0.0927	82.76	77.03	-1.69	81.41
	SVR	0.0650	84.02	78.19	-14.34	67.44
	RF	0.0561	84.41	78.56	+0.90	85.33

The exclusion of ALOS PALSAR-2 (L-band) caused a dramatic reduction in predictive power, with the best-performing models dropping from an of ~0.34 (Full Workflow) to ~0.18 (Sentinel-Only).

1. Failure of Filter Methods (MI & SelectKBest): The models utilizing features selected by Mutual Information and SelectKBest performed poorly, with values collapsing to < 0.11. These methods selected only optical indices. The extremely low accuracy (for RF) confirms that optical data alone—even with sophisticated indices like NBR and NDRE—is statistically insufficient for predicting tropical biomass due to rapid signal saturation.
2. Resilience of PCA and RFE: The PCA and RFE methods demonstrated higher resilience, achieving values between 0.17 – 0.18.
   - PCA achieved the single highest accuracy (with MLR), suggesting that condensing the variance of the entire spectral and texture dataset was the most effective way to squeeze information from the limited Sentinel sensors.
   - RFE followed closely (with MLP), validating its selection strategy. By forcing the inclusion of C-band texture (VH_contrast) and topography, RFE managed to salvage some structural predictive power that the optical-only models missed.
3. Model Behavior: Interestingly, in this low-information environment, the complex non-linear models (RF, SVR) often performed worse than the simple linear baseline (MLR). For instance, with RFE features, the Random Forest dropped to while MLR maintained . This indicates that when the primary driver of the relationship (L-band volume scattering) is missing, complex models may overfit to noise, whereas linear regression captures the broad, albeit weak, trend between vegetation greenness and biomass.

4.4.4 Best Performing Model Without L-Band SAR

Based on the comparative evaluation of the Sentinel-only feature space, the combination of Principal Component Analysis (PCA) feature selection with Multiple Linear Regression (MLR) emerged as the most accurate configuration.

This model achieved an R^2 of 0.1807 and an RMSE of 78.65 $Mg \cdot ha^{-1}$.

It is notable that in this restricted scenario, the complex non-linear models (RF, SVR, MLP) failed to outperform the linear baseline. This suggests that the relationship between optical/C-band inputs and biomass is dominated by weak linear correlations (e.g., greenness vs. biomass) rather than the complex structural interactions that Deep Learning models excel at extracting from L-band data.

To quantify the specific impact of excluding ALOS PALSAR-2, the performance of this best Sentinel-only model (PCA + MLR) was compared directly against the overall best model from the full workflow (RFE + MLP, Section 4.2).

Table 4.3 Performance degradation analysis due to L-band exclusion

Scenario	Model Configuration	R^2	RMSE (Mg/ha)	Bias (Mg/ha)	MAPE (%)
Full Fusion (With L-Band)	**RFE + MLP**	**0.3432**	**74.02**	**-1.26**	**69.98**
Sentinel-Only (No L-Band)	**PCA + MLR**	**0.1807**	**78.65**	**-1.51**	**73.63**
Performance Drop		**-47.3%**	**+4.63**	**-0.25**	**+3.65%**

The exclusion of L-band SAR data resulted in a critical degradation of predictive performance, as detailed in Table 4.3. The model's explanatory power (R^2) collapsed by approximately 47% (from 0.3432 to 0.1807), confirming that nearly half of the variance explained by the full model was driven specifically by the L-band interaction with woody biomass.

A significant finding in this restricted experiment is the shift in the optimal algorithm. In the full multi-sensor workflow, the non-linear Multi-Layer Perceptron (MLP) was superior (R^2=0.34), effectively modeling the complex volume scattering provided by ALOS PALSAR-2. However, in the Sentinel-only scenario, the simple Multiple Linear Regression (MLR) emerged as the top performer (R^2=0.1807), marginally outperforming the MLP (R^2=0.1774).

This shift indicates that without the structural depth of L-band radar, the relationship between satellite inputs and biomass simplifies to a weak linear trend dominated by optical greenness (chlorophyll). Advanced deep learning models (MLP) failed to find additional patterns because the C-band and Optical data saturate too early, effectively rendering the "deep" architecture unnecessary. Furthermore, the systematic negative bias worsened across most models, with SVR exhibiting a severe underestimation bias of -12.76 Mg/ha, proving its inability to capture high-biomass forest stocks without L-band inputs.

4.5 Final Inference Using the Best Model

Following the rigorous validation and testing phases, the Multi-Layer Perceptron (MLP) model, utilizing the feature subset selected by Recursive Feature Elimination (RFE), was identified as the optimal architecture for spatial prediction. This section details the application of this best-performing model to generate a wall-to-wall carbon stock map for the entire study area.

4.5.1 Generation of the Spatial Carbon Stock Map

The final inference process involved applying the trained MLP network to the complete multi-sensor image stack prepared for the year 2020. This stack consisted of the 15 specific predictors identified by the RFE analysis, comprising a fusion of:

- Optical: Sentinel-2 bands and indices (NDRE, NBR) for vegetation health.
- SAR: Sentinel-1 (C-band) and ALOS PALSAR (L-band) backscatter for structural volume.
- Topography: SRTM-derived elevation and slope.

The model processed the study area pixel-by-pixel at a spatial resolution of 10 meters. For every valid pixel grid location, the MLP calculated a predicted Above-Ground Biomass (AGB) value, which was subsequently converted to Carbon Stock ($Mg \cdot ha^{-1}$) using the conversion factor of 0.47. To ensure physical validity, non-forest areas (water bodies, built-up land) were masked out using the ESA WorldCover dataset, and any negative predictions resulting from regression artifacts were clipped to zero.

4.5.2 Spatial Distribution Analysis

The resulting spatial distribution of forest carbon stocks for the Special Region of Yogyakarta is presented in Figure 4.5.

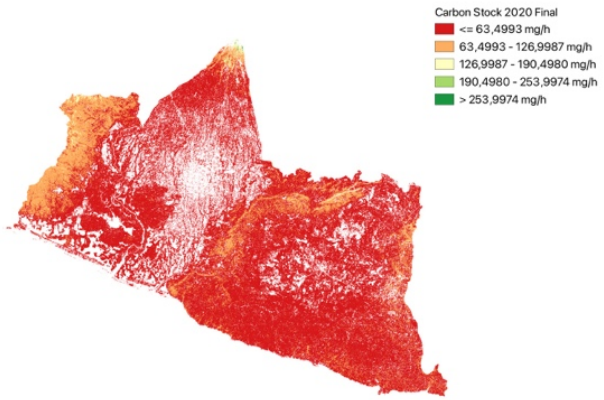


Figure 4.5 illustrate the predicted carbon stock distribution for the year 2020 using the MLP.

The map reveals distinct spatial heterogeneity in carbon storage across the region, aligning with known land cover gradients:

- High Carbon Stock Areas (>140 Mg C/ha): In the generated spatial map, regions displayed in white or bright colors represent biomass densities exceeding 140 Mg/ha. Physically, these areas correspond to the dense, mature forests of Mount Merapi National Park.
- Moderate Carbon Stock Areas (95 – 190 Mg C/ha): The yellow to light green transition zones typically indicate secondary forests, agroforestry systems (mixed gardens), or plantation forests (e.g., teak or pine) which are common in the peripheral rural areas of Yogyakarta.
- Low Carbon Stock Areas (< 63 Mg C/ha): Depicted in red, these areas likely represent degraded lands, young regrowth, or sparsely vegetated areas on the fringes of human settlements.

The successful generation of this high-resolution map demonstrates the capability of the MLP + RFE framework to upscale point-based GEDI estimates into a continuous regional baseline, providing critical data for carbon accounting and forest management.

However, it is critical to note that values plateauing in this 'white' range may indicate sensor saturation. While the inclusion of L-band SAR extended the sensitivity beyond the typical 100 Mg/ha limit of optical sensors, the model still exhibits asymptotic behavior at extremely high densities (>400 Mg/ha), effectively 'capping' the predictions. Thus, these bright regions should be interpreted as 'High Conservation Value Forest' with high certainty, though

the exact biomass value may be underestimated due to the physical limitations of the saturation point.

4.6 Discussion

4.6.1 Efficacy of Feature Selection Strategies

The comparison of feature selection methods in this study revealed that Recursive Feature Elimination (RFE) provided the most ecologically robust predictor set. Unlike filter methods (Mutual Information and SelectKBest), which evaluate features in isolation, RFE utilizes the interaction within the model to identify synergies between variables. This was evident in the selection of L-band (ALOS PALSAR) and C-band (Sentinel-1) backscatter alongside Optical (Sentinel-2) indices. Mutual Information, conversely, exhibited a strong bias toward optical variables (retaining redundant TCI bands), likely due to the high signal-to-noise ratio of optical data compared to the inherent speckle noise of SAR. Consequently, models using MI inputs lacked the structural information required to resolve high-biomass saturation.

Principal Component Analysis (PCA) also performed exceptionally well, matching the accuracy of RFE. This suggests that while individual bands may contain noise (e.g., atmospheric artifacts in Sentinel-2 or speckle in Sentinel-1), projecting them into orthogonal components effectively concentrates the information signal and reduces noise. However, the trade-off is a complete loss of physical interpretability, making RFE the preferred method for explaining which biophysical parameters drive carbon storage.

4.6.2 Neural Networks vs. Traditional Regression

The superior performance of the Multi-Layer Perceptron (MLP) highlights the necessity of non-linear modeling in tropical biomass estimation. Satellite signals—particularly from optical and C-band sensors—exhibit an asymptotic relationship with biomass, saturating at approximately 100–150 Mg/ha.

- Linear Models (MLR): The Multiple Linear Regression model yielded the lowest accuracy because it forces a straight-line fit onto this curved relationship, resulting in massive underestimation of high-biomass areas and overestimation of low-biomass areas.
- Artificial Neural Networks (MLP): The MLP, with its hidden layers and non-linear activation functions, successfully approximated this complex saturation curve. Unlike the rigid linear structure of MLR, the neural network architecture effectively "learned"

to weigh the optical inputs for low-biomass vegetation and shift reliance to the deeper-penetrating L-band radar features for high-biomass forests.

4.6.3 The Role of Multi-Sensor Fusion

The results confirm that no single sensor is sufficient for comprehensive carbon mapping. Optical sensors (Sentinel-2) excelled at delineating vegetation extent and health (chlorophyll) but saturated quickly. C-band SAR (Sentinel-1) provided sensitivity to canopy roughness but struggled with volume scattering in dense forests. The integration of ALOS PALSAR (L-band) was the critical differentiator; its longer wavelength (approx. 23 cm) penetrated the canopy to interact with trunks and branches, providing the necessary volume scattering information to push predictions beyond the optical saturation point. The RFE + MLP model's success is a direct result of effectively fusing these complementary physical measurements.

4.6.4 Uncertainties and the Impact of GEDI Noise

While the models performed well, the scatter plots (Section 4.3) revealed significant heteroscedasticity, with prediction errors increasing in high-biomass regions. This uncertainty is not solely due to model limitations but is also attributable to the GEDI LiDAR training data itself. GEDI footprints (25m diameter) have inherent geolocation uncertainties (approx. 10m) and waveform processing errors, particularly on steep slopes common in Yogyakarta's mountainous regions. Consequently, the "ground truth" used for training contains noise. The MLP model's ability to achieve an R^2 of 0.34 despite this noisy training data indicates a high degree of robustness, but it also suggests that future improvements may require filtering GEDI shots more aggressively for slope quality or incorporating airborne LiDAR for calibration.

CHAPTER V
CONCLUSION

5.1 Conclusion

This study evaluated the efficacy of multi-sensor data fusion—integrating Sentinel-1, Sentinel-2, and ALOS PALSAR-2—for mapping tropical forest carbon stocks in the Special Region of Yogyakarta. By testing sixteen experimental configurations, the research identified the optimal modeling strategy while exposing critical limitations in standard approaches.

The integration of Recursive Feature Elimination (RFE) with the Multi-Layer Perceptron (MLP) emerged as the most robust architecture. This combination achieved the highest accuracy with an R^2 of 0.3432, RMSE of 74.02 Mg/ha, and MAPE of 69.98%. The MLP successfully modeled non-linear vegetation dynamics that traditional linear models failed to resolve, effectively leveraging the synergistic information from optical and radar sensors.

While the Support Vector Regression (SVR) model yielded the lowest percentage error (MAPE 54.41%), it was found scientifically unsuitable for operational reporting due to a significant systematic negative bias (-14.25 Mg/ha). This bias indicates that SVR consistently underestimates biomass in high-density forests (>140 Mg/ha), effectively "capping" predictions and failing to detect critical carbon sinks.

Ineffectiveness of Filter Methods: Univariate feature selection methods (Mutual Information and SelectKBest) consistently underperformed ($R^2 < 0.26$). These methods failed because they prioritized optical indices (e.g., NDVI) which saturate early, while discarding the noisy but essential L-band radar backscatter required to penetrate dense canopies.

The study confirms that removing L-band data causes model performance to collapse (RMSE increases by +4.64 Mg/ha). Optical and C-band sensors alone are insufficient for tropical biomass mapping due to signal saturation at ~100 Mg/ha. The inclusion of L-band volume scattering is therefore a strict requirement for accurate national carbon reporting.

5.2 Limitations of the Study

Despite the robust results, three key limitations must be acknowledged:

- Reference Data Uncertainty: The model relies on GEDI Level 4A products as "ground truth." Since GEDI itself is a modeled estimate with geolocation errors (~10m), any inherent biases in the reference data are propagated into the final predictions.

- Saturation at Extreme Values: Although multi-sensor fusion delayed saturation, the model still underestimates biomass in extremely dense forests (>400 Mg/ha). This suggests a physical limit to the sensitivity of the current satellite sensor suite.
- Temporal Constraints: The analysis is restricted to a single year (2020), meaning it does not account for seasonal phenology or long-term forest degradation trends.

5.3 Recommendations

To advance the operational readiness of this framework for Indonesia's REDD+ MRV (Measurement, Reporting, and Verification) systems, the following future works are recommended:

1. Hybrid Calibration: Integrate direct National Forest Inventory (NFI) field plots to calibrate GEDI samples locally before training, thereby reducing error propagation.
2. Spatial Deep Learning: Move beyond pixel-based approaches by implementing Convolutional Neural Networks (CNNs) to exploit spatial texture patterns in radar imagery.
3. Uncertainty Mapping: Implement Quantile Regression to generate pixel-wise confidence intervals, providing policymakers with a reliability metric for every hectare of mapped forest.

REFERENCE

Abbas, S., Wong, M. S., Wu, J., Shahzad, N., & Irteza, S. M. (2020). Approaches of satellite remote sensing for the assessment of above-ground biomass across tropical forests: Pan-tropical to national scales. In *Remote Sensing* (Vol. 12, Issue 20, pp. 1–38). MDPI AG. https://doi.org/10.3390/rs12203351

Abdumalikov, S., Kim, J., & Yoon, Y. (2024). Performance Analysis and Improvement of Machine Learning with Various Feature Selection Methods for EEG-Based Emotion Classification. *Applied Sciences 2024, Vol. 14, Page 10511*, *14*(22), 10511. https://doi.org/10.3390/APP142210511

Adamu, B., Ibrahim, S., Rasul, A., Whanda, S. J., Headboy, P., Muhammed, I., & Maiha, I. A. (2021). Evaluating the accuracy of spectral indices from Sentinel-2 data for estimating forest biomass in urban areas of the tropical savanna. *Remote Sensing Applications: Society and Environment*, *22*, 100484. https://doi.org/10.1016/J.RSASE.2021.100484

Antunes, R., Junior, L., Costa, G., Feitosa, R., de Souza Bias, E., Cereda Junior, A., Almeida, C., Cué La Rosa, L. E., Happ, P., & Chiamulera, L. (2024). Leveraging SAR and Optical Remote Sensing for Enhanced Biomass Estimation in the Amazon with Random Forest and XGBoost Models. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *10*(3), 21–27. https://doi.org/10.5194/isprs-annals-X-3-2024-21-2024

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

Butler, B. J., Sass, E. M., Gamarra, J. G. P., Campbell, J. L., Wayson, C., Olguín, M., Carrillo, O., & Yanai, R. D. (2024). Uncertainty in REDD+ carbon accounting: a survey of experts involved in REDD+ reporting. *Carbon Balance and Management*, *19*(1). https://doi.org/10.1186/s13021-024-00267-z

Chen, Z., Yang, X., Pan, X., Wu, T., Lei, J., Chen, X., Li, Y., & Chen, Y. (2025). Estimating Forest Aboveground Biomass in Tropical Zones by Integrating LiDAR and Sentinel-2B Data. *Sustainability (Switzerland)*, *17*(8). https://doi.org/10.3390/su17083631

Cheng, X. (2024). A Comprehensive Study of Feature Selection Techniques in Machine Learning Models. *Insights in Computer, Signals and Systems*, *1*(1). https://soapubs.com/index.php/ICSS

David, R. M., Rosser, N. J., & Donoghue, D. N. M. (2022). Improving above ground biomass estimates of Southern Africa dryland forests by combining Sentinel-1 SAR and Sentinel-2 multispectral imagery. *Remote Sensing of Environment*, *282*, 113232. https://doi.org/10.1016/J.RSE.2022.113232

Delloye, C., Weiss, M., & Defourny, P. (2018). Retrieval of the canopy chlorophyll content from Sentinel-2 spectral bands to estimate nitrogen uptake in intensive winter wheat cropping systems. *Remote Sensing of Environment*, *216*, 245–261. https://doi.org/10.1016/J.RSE.2018.06.037

Doraisami, M., Domke, G. M., & Martin, A. R. (2024). Improving wood carbon fractions for multiscale forest carbon estimation. In *Carbon Balance and Management* (Vol. 19, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13021-024-00272-2

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, *120*, 25–36. https://doi.org/10.1016/J.RSE.2011.11.026

Dubayah, R. O., Armston, J., Kellner, J. R., Duncanson, L., Healey, S. P., Patterson, P. L., Hancock, S., Tang, H., Hofton, M. A., Blair, J. B., & Luthcke, S. B. (2022). *GEDI L4A Footprint Level Aboveground Biomass Density, Golden Weeks, Version 1*. NASA ORNL DAAC. https://doi.org/10.3334/ORNLDAAC/2028

Ferdinandus Edwin Penalun, Arief Hermawan, Donny Avianto, & Arif Pramudwiatmoko. (2024). A Multi-Layer Perceptron Regression and Variant Windowing for Estimating Rainfall Based on Weather Radar Data. *Journal of Education and Science*, *33*(2), 58–71. https://doi.org/10.33899/edusj.2024.146355.1421

Food and Agriculture Organization of the United Nations. (2010). *Global Forest Resources Assessment 2010: Terms and definitions*. www.fao.org/forestry/fra

Hoffmann, H. (n.d.). *Kernel PCA for Novelty Detection*.

Kelsey, K. C., & Neff, J. C. (2014). Estimates of aboveground biomass from texture analysis of landsat imagery. *Remote Sensing*, *6*(7), 6407–6422. https://doi.org/10.3390/rs6076407

Khan, K., Iqbal, J., Ali, A., & Khan, S. N. (2020). Assessment of sentinel-2-derived vegetation indices for the estimation of above-ground biomass/carbon stock, temporal deforestation and carbon emissions estimation in the moist temperate forests of pakistan. *Applied Ecology and Environmental Research*, *18*(1), 783–815. https://doi.org/10.15666/aeer/1801_783815

Liu, A., Chen, Y., & Cheng, X. (2025). Improving Tropical Forest Canopy Height Mapping by Fusion of Sentinel-1/2 and Bias-Corrected ICESat-2–GEDI Data. *Remote Sensing*, *17*(12). https://doi.org/10.3390/rs17121968

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, *28*(5), 823–870. https://doi.org/10.1080/01431160600746456;REQUESTEDJOURNAL:JOURNAL:TRES20;ISSUE:ISSUE:DOI

Ma, C., Li, X., & McCabe, M. F. (2020). Retrieval of High-Resolution Soil Moisture through Combination of Sentinel-1 and Sentinel-2 Data. *Remote Sensing 2020, Vol. 12, Page 2303*, *12*(14), 2303. https://doi.org/10.3390/RS12142303

Mansingh, A., Pradhan, A., Sahoo, S. R., Cherwa, S. S., Mishra, B. P., Rath, L. P., Ekka, N. J., & Panda, B. P. (2025). Tree diversity, population structure, biomass accumulation, and carbon stock dynamics in tropical dry deciduous forests of Eastern India. *BMC Ecology and Evolution*, *25*(1). https://doi.org/10.1186/s12862-025-02385-9

Migolet, P., Goïta, K., Pambo, A. F. K., & Mambimba, A. N. (2022). Estimation of the total dry aboveground biomass in the tropical forests of Congo Basin using optical, LiDAR, and radar data. *GIScience & Remote Sensing*, *59*(1), 431–460. https://doi.org/10.1080/15481603.2022.2026636

Mo, L., Zohner, C. M., Reich, P. B., Liang, J., de Miguel, S., Nabuurs, G. J., Renner, S. S., van den Hoogen, J., Araza, A., Herold, M., Mirzagholi, L., Ma, H., Averill, C., Phillips, O. L., Gamarra, J. G. P., Hordijk, I., Routh, D., Abegg, M., Adou Yao, Y. C., … Crowther, T. W. (2023). Integrated global assessment of the natural forest carbon potential. *Nature*, *624*(7990), 92–101. https://doi.org/10.1038/s41586-023-06723-z

Musthafa, M., & Singh, G. (2022). Improving Forest Above-Ground Biomass Retrieval Using Multi-Sensor L- and C- Band SAR Data and Multi-Temporal Spaceborne LiDAR Data. *Frontiers in Forests and Global Change*, *5*. https://doi.org/10.3389/ffgc.2022.822704

Ni-Meister, W., Rojas, A., & Grant, I. (2025). Enhancing aboveground biomass density estimates in tropical forests using GEDI waveform data. *Science of Remote Sensing*, *12*, 100286. https://doi.org/10.1016/J.SRS.2025.100286

Opelele, O. M., Yu, Y., Fan, W., Chen, C., & Kachaka, S. K. (2021). Biomass estimation based on multilinear regression and machine learning algorithms in the mayombe tropical forest, in the democratic republic of congo. *Applied Ecology and Environmental Research*, *19*(1), 359–377. https://doi.org/10.15666/aeer/1901_359377

Papaioannou, N., Myllis, G., Tsimpiris, A., & Vrana, V. (2025). The Role of Mutual Information Estimator Choice in Feature Selection: An Empirical Study on mRMR. *Information 2025, Vol. 16, Page 724*, *16*(9), 724. https://doi.org/10.3390/INFO16090724

Prada, C. M., Heineman, K. D., Pardo, M. J., Piponiot, C., & Dalling, J. W. (2025). Soil and biomass carbon storage is much higher in Central American than Andean montane forests. *Biogeosciences*, *22*(14), 3615–3634. https://doi.org/10.5194/bg-22-3615-2025

Praputra, A. V., Bong, I. W., Ekowati, D., Hofstee, C., & Maryudi, A. (2016). Getting the data flowing: Lessons learned from existing reporting systems in the forestry sector in Indonesia for REDD+ MRV. *PLoS ONE*, *11*(11). https://doi.org/10.1371/journal.pone.0156743

Pritalia, G. L. (2022). *Analisis Komparatif Algoritme Machine Learning pada Klasifikasi Kualitas Air Layak Minum* (Vol. 2, Issue 1).

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. In *Frontiers in Bioinformatics* (Vol. 2). Frontiers Media SA. https://doi.org/10.3389/fbinf.2022.927312

Rodríguez-Pérez, R., & Bajorath, J. (2022). Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *Journal of Computer-Aided Molecular Design*, *36*(5), 355–362. https://doi.org/10.1007/s10822-022-00442-9

Rosenqvist, A., Shimada, M., Suzuki, S., Ohgushi, F., Tadono, T., Watanabe, M., Tsuzuku, K., Watanabe, T., Kamijo, S., & Aoki, E. (2014). Operational performance of the ALOS global systematic acquisition strategy and observation plans for ALOS-2 PALSAR-2. *Remote Sensing of Environment*, *155*, 3–12. https://doi.org/10.1016/J.RSE.2014.04.011

Singh, A., Kushwaha, S., Alarfaj, M., & Singh, M. (2022). Comprehensive Overview of Backpropagation Algorithm for Digital Image Denoising. *Electronics (Switzerland)*, *11*(10). https://doi.org/10.3390/electronics11101590

Slagter, B., Reiche, J., Marcos, D., Mullissa, A., Lossou, E., Peña-Claros, M., & Herold, M. (2023). Monitoring direct drivers of small-scale tropical forest disturbance in near real-time with Sentinel-1 and -2 data. *Remote Sensing of Environment*, *295*, 113655. https://doi.org/10.1016/J.RSE.2023.113655

Soto-Navarro, C., Ravilious, C., Arnell, A., De Lamo, X., Harfoot, M., Hill, S. L. L., Wearn, O. R., Santoro, M., Bouvet, A., Mermoz, S., Le Toan, T., Xia, J., Liu, S., Yuan, W., Spawn, S. A., Gibbs, H. K., Ferrier, S., Harwood, T., Alkemade, R., … Kapos, V. (2020). Mapping co-benefits for carbon storage and biodiversity to inform conservation policy and action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1794). https://doi.org/10.1098/rstb.2019.0128

Tello, M., Cazcarra-Bes, V., Pardini, M., & Papathanassiou, K. (2018). Forest structure characterization from SAR tomography at L-band. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(10), 3402–3414. https://doi.org/10.1109/JSTARS.2018.2859050

Tian, L., Wu, X., Tao, Y., Li, M., Qian, C., Liao, L., & Fu, W. (2023). Review of Remote Sensing-Based Methods for Forest Aboveground Biomass Estimation: Progress, Challenges, and Prospects. *Forests 2023, Vol. 14, Page 1086*, *14*(6), 1086. https://doi.org/10.3390/F14061086

Vapnik, V. N. (1995). *The Nature of Statistical Learning 2010*. Springer.

Yang, H., Song, M., Son, H., Kim, R., & Choi, E. (2025). Evaluating REDD+ Readiness: High-Potential Countries Based on MRV Capacity. *Forests*, *16*(1). https://doi.org/10.3390/f16010067

Yu, Y., Lei, Y., Siqueira, P., Liu, X., Gu, D., Fu, A., Pang, Y., Huang, W., & Shi, J. (2025). Large-scale forest stand height mapping in the northeastern US and China using L-band spaceborne repeat-pass InSAR and GEDI lidar data. *Earth System Science Data*, *17*(9), 4397–4429. https://doi.org/10.5194/essd-17-4397-2025

Yu, Y., & Saatchi, S. (2016). Sensitivity of L-band SAR backscatter to aboveground biomass of global forests. *Remote Sensing*, *8*(6). https://doi.org/10.3390/rs8060522

Zhang, Y., Liu, J., & Shen, W. (2022). A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. In *Applied Sciences (Switzerland)* (Vol. 12, Issue 17). MDPI. https://doi.org/10.3390/app12178654

APPENDIX

Appendix A: Source Code

A.1 Data Acquisition Script

```
import ee
import geopandas as gpd
import json
import os
import time
# === 1. USER SETTINGS ===
GEOJSON_PATH = "yogyakarta_boundary.geojson"
GEE_PROJECT_ID = "skripsi-475914"
YEAR = '2020'
DRY_SEASON_START = f'{YEAR}-06-01'
DRY_SEASON_END = f'{YEAR}-09-30'
TARGET_RESOLUTION = 10
OUTPUT_FILE_NAME = f"All_Rasters_Yogyakarta_{YEAR}"
OUTPUT_DRIVE_FOLDER = "raster_exports"
# === 2. AUTHENTICATE GEE ===
try:
ee.Initialize(project=GEE_PROJECT_ID)
print("✅ Earth Engine initialized successfully.")
except Exception as e:
print(f"❌ Error initializing GEE: {e}")
raise
# === 3. LOAD AOI ===
print(f"📂 Loading AOI: {GEOJSON_PATH}")
if not os.path.exists(GEOJSON_PATH):
print(f"❌ ERROR: Cannot find file '{GEOJSON_PATH}'")
exit()
gdf = gpd.read_file(GEOJSON_PATH).to_crs(epsg=4326)
geom_json = json.loads(gdf.geometry.to_json())
if len(geom_json['features']) == 1:
coords = geom_json['features'][0]['geometry']['coordinates']
geom_type = geom_json['features'][0]['geometry']['type']
aoi = ee.Geometry.MultiPolygon(coords) if geom_type == 'MultiPolygon' else
ee.Geometry.Polygon(coords)
else:
geoms = [
ee.Geometry.Polygon(f['geometry']['coordinates'])
if f['geometry']['type'] == 'Polygon'
else ee.Geometry.MultiPolygon(f['geometry']['coordinates'])
for f in geom_json['features']
]
aoi = ee.Geometry.MultiPolygon(geoms).union()
print("✅ AOI loaded successfully.")
# === 4. PROCESS SENTINEL-2 (OPTICAL) ===
print("🛰️ Processing Sentinel-2 (Optical)...")
def maskS2clouds(image):
qa = image.select('QA60')
cloudBitMask = 1 << 10
cirrusBitMask = 1 << 11
mask = qa.bitwiseAnd(cloudBitMask).eq(0).And(qa.bitwiseAnd(cirrusBitMask).eq(0))
return image.updateMask(mask).divide(10000).toFloat()
s2_collection = (ee.ImageCollection('COPERNICUS/S2_SR_HARMONIZED')
.filterBounds(aoi)
.filterDate(DRY_SEASON_START, DRY_SEASON_END)
.filter(ee.Filter.lt('CLOUDY_PIXEL_PERCENTAGE', 20))
.map(maskS2clouds))
s2_composite = s2_collection.median().clip(aoi)
```

```
nir = s2_composite.select('B8')
red = s2_composite.select('B4')
blue = s2_composite.select('B2')
green = s2_composite.select('B3')
rededge1 = s2_composite.select('B5')
rededge2 = s2_composite.select('B8A')
swir1 = s2_composite.select('B11')
swir2 = s2_composite.select('B12')
ndvi = nir.subtract(red).divide(nir.add(red)).rename('NDVI').toFloat()
evi = s2_composite.expression('2.5 * ((NIR - RED) / (NIR + 6 * RED - 7.5 * BLUE +
1))', {'NIR': nir, 'RED': red, 'BLUE': blue}).rename('EVI').toFloat()
savi = s2_composite.expression('((NIR - RED) / (NIR + RED + 0.5)) * 1.5', {'NIR':
nir, 'RED': red}).rename('SAVI').toFloat()
msavi2 = s2_composite.expression('(2 * NIR + 1 - sqrt((2 * NIR + 1)**2 - 8 * (NIR -
RED))) / 2', {'NIR': nir, 'RED': red}).rename('MSAVI2').toFloat()
ndmi = nir.subtract(swir1).divide(nir.add(swir1)).rename('NDMI').toFloat()
nbr = nir.subtract(swir2).divide(nir.add(swir2)).rename('NBR').toFloat()
gndvi = nir.subtract(green).divide(nir.add(green)).rename('GNDVI').toFloat()
ndre =
rededge2.subtract(rededge1).divide(rededge2.add(rededge1)).rename('NDRE').toFloat()
s2_final = s2_composite.addBands([ndvi, evi, savi, msavi2, ndmi, nbr, gndvi, ndre])
print("✅ Sentinel-2 processing complete.")
s2_projection = s2_final.select('B2').projection()
# === 5. PROCESS SENTINEL-1 (RADAR + TEXTURE) ===
print("🛰️ Processing Sentinel-1 (C-Band Radar + Texture)...")
s1 = (ee.ImageCollection('COPERNICUS/S1_GRD')
.filterBounds(aoi)
.filterDate(DRY_SEASON_START, DRY_SEASON_END)
.filter(ee.Filter.eq('instrumentMode', 'IW'))
.filter(ee.Filter.listContains('transmitterReceiverPolarisation', 'VV'))
.filter(ee.Filter.listContains('transmitterReceiverPolarisation', 'VH')))
def apply_speckle_filter(image):
vv = image.select('VV').focal_median(1)
hv = image.select('VH').focal_median(1)
return image.addBands(vv, None, True).addBands(hv, None, True)
s1_filtered = s1.map(apply_speckle_filter)
s1_composite = s1_filtered.median().clip(aoi)
print("Calculating S1 Radar Texture (GLCM)...")
vh_8bit = s1_composite.select('VH').unitScale(-25, 0).multiply(255).toInt8()
glcm = vh_8bit.glcmTexture(size=1)
texture_bands = glcm.select('VH_asm', 'VH_contrast', 'VH_corr', 'VH_var', 'VH_idm',
'VH_savg', 'VH_ent').toFloat()
s1_final = s1_composite.select(['VV', 'VH']).addBands(texture_bands).toFloat()
print("✅ Sentinel-1 processing complete.")
# === 6. PROCESS SRTM (TOPOGRAPHY) ===
print("🛰️ Processing SRTM (Topography)...")
srtm = ee.Image("USGS/SRTMGL1_003").clip(aoi)
elevation = srtm.select('elevation')
slope = ee.Terrain.slope(elevation).rename('slope')
aspect = ee.Terrain.aspect(elevation).rename('aspect')
srtm_bands = elevation.addBands(slope).addBands(aspect).toFloat()
print(" Resampling SRTM from 30m to 10m grid...")
srtm_final = srtm_bands.resample('bilinear').reproject(
crs=s2_projection,
scale=TARGET_RESOLUTION
)
print("✅ SRTM processing complete.")
# === 7. PROCESS ALOS PALSAR-1 (L-BAND RADAR) ===
print("🛰️ Processing ALOS PALSAR-1 (L-Band Radar)...")
alos = (ee.ImageCollection('JAXA/ALOS/PALSAR/YEARLY/SAR')
.filterDate(f'{YEAR}-01-01', f'{YEAR}-12-31')
.filterBounds(aoi))
alos_image = alos.first().clip(aoi)
hv = alos_image.select('HV')
hh = alos_image.select('HH')
ratio = hv.divide(hh.add(0.0001)).rename('HV_HH_Ratio')
```

```
alos_bands = alos_image.select(['HH', 'HV']).addBands(ratio).toFloat()
print(" Resampling ALOS PALSAR from 25m to 10m grid...")
alos_final = alos_bands.resample('bilinear').reproject(
crs=s2_projection,
scale=TARGET_RESOLUTION
)
print("✅ ALOS PALSAR processing complete.")
# === 8. COMBINE ALL BANDS ===
print("Combining all rasters into one image...")
master_image =
s2_final.addBands(s1_final).addBands(srtm_final).addBands(alos_final)
print(f"✅ Master image created with {len(master_image.bandNames().getInfo())}
bands.")
# === 9. EXPORT TO GOOGLE DRIVE ===
print(f"📤 Exporting to Google Drive as '{OUTPUT_FILE_NAME}.tif'...")
task = ee.batch.Export.image.toDrive(
image=master_image,
description=OUTPUT_FILE_NAME,
folder=OUTPUT_DRIVE_FOLDER,
fileNamePrefix=OUTPUT_FILE_NAME,
region=aoi,
scale=TARGET_RESOLUTION, # Export at 10m
maxPixels=1e13
)
task.start()
print("\n🚀 Export started! Check progress on: https://code.earthengine.google.com
-> Tasks tab")
print(" This is a VERY HEAVY task and may take > 1 hour. The script will wait...")
# === 10. (OPTIONAL) MONITOR STATUS ===
while task.active():
print(f"Task state: {task.status()['state']}... (waiting 30s)")
time.sleep(30)
status = task.status()
print(f"✅ Task finished with state: {status['state']}")
if 'error_message' in status:
print(f"❌ ERROR: {status['error_message']}")
else:
print(f"🎉 Success! Check your '{OUTPUT_DRIVE_FOLDER}' folder in Google Drive.")
```

A.2 Extract GEDI L4A to CSV

```
import h5py
import geopandas as gpd
import pandas as pd
import numpy as np
import glob
import os
import re
from shapely.geometry import Point
from tqdm import tqdm
from datetime import datetime
# === 1. DEFINE INPUT AND OUTPUT FILES ===
BASE_DIR = os.path.dirname(os.path.abspath(__file__))
# --- INPUTS ---
# 1. Your local GeoJSON boundary file
aoi_path = os.path.join(BASE_DIR, "yogyakarta_boundary.geojson")
# 2. Folder where you downloaded the GEDI L4A granules for 2020
```

```
# *** This folder MUST contain GEDI L4A files ***
h5_folder_path = os.path.join(BASE_DIR, "Gedi L4A 2020", "*.h5")
# --- OUTPUT ---
# 3. The new CSV with AGBD (carbon stock) for 2020
output_csv = os.path.join(BASE_DIR, "gedi_l4a_2020_filtered.csv")
# === 2. LOAD THE AREA OF INTEREST (AOI) ===
print(f"📍 Loading AOI from: {aoi_path}")
try:
aoi = gpd.read_file(aoi_path).to_crs("EPSG:4326")
if aoi.empty:
raise ValueError("AOI file is empty.")
print("✅ AOI loaded successfully.")
except Exception as e:
print(f"❌ Error loading AOI file: {e}")
exit()
# === 3. FIND GEDI L4A FILES ===
h5_files = glob.glob(h5_folder_path)
if not h5_files:
print(f"❌ No .h5 files found in '{os.path.join(BASE_DIR, 'Gedi L4A 2020')}'")
print(" Please download the GEDI L4A granules for 2020 from NASA Earthdata.")
exit()
print(f"🛰️ Found {len(h5_files)} GEDI L4A H5 files. Starting processing...\n")
records = []
total_points_processed = 0
total_points_kept = 0
# === 4. LOOP OVER ALL GEDI FILES ===
for file_path in tqdm(h5_files, desc="Processing GEDI L4A files"):
file_name = os.path.basename(file_path)
# --- Extract time info from filename ---
match = re.search(r"_(\d{13})_", file_name) # Match YYYYJJJHHMMSS
year, month = None, None
if match:
datestring = match.group(1)
try:
dt = datetime.strptime(datestring[:7], "%Y%j")
year = dt.year
month = dt.month
except Exception:
month = None
try:
with h5py.File(file_path, "r") as f:
beam_keys = [key for key in f.keys() if key.startswith('BEAM')]
if not beam_keys:
tqdm.write(f"⚠️ No BEAMs found in {file_name}. Skipping.")
```

```
        continue
        file_lats, file_lons, file_agbds = [], [], []
        file_l4_flags, file_degrade_flags = [], [] # To store quality flags
        for beam in beam_keys:
        try:
        beam_group = f[beam]
        # --- Extract all necessary L4A data (AGBD) ---
        agbd = beam_group['agbd'][:]
        lat = beam_group['lat_lowestmode'][:]
        lon = beam_group['lon_lowestmode'][:]
        l4_flag = beam_group['l4_quality_flag'][:]
        degrade_flag = beam_group['degrade_flag'][:]
        if not (len(agbd) == len(lat) == len(lon) == len(l4_flag) == len(degrade_flag)):
        tqdm.write(f"⚠️ Beam {beam} in {file_name} has mismatched data lengths. Skipping
beam.")
        continue
        file_lats.append(lat)
        file_lons.append(lon)
        file_agbds.append(agbd)
        file_l4_flags.append(l4_flag)
        file_degrade_flags.append(degrade_flag)
        except KeyError as ke:
        tqdm.write(f"ℹ️ Beam {beam} in {file_name} missing key {ke}. Skipping beam.")
        except Exception as e:
        tqdm.write(f"ℹ️ Beam {beam} in {file_name} had error: {e}. Skipping beam.")
        if not file_lats: # If no beams had valid data
        tqdm.write(f"⚠️ No valid beam data in {file_name}. Skipping.")
        continue
        # --- Flatten all data from all beams into 1D arrays ---
        lat_flat = np.concatenate(file_lats).flatten()
        lon_flat = np.concatenate(file_lons).flatten()
        agb_flat = np.concatenate(file_agbds).flatten()
        l4_flag_flat = np.concatenate(file_l4_flags).flatten()
        degrade_flag_flat = np.concatenate(file_degrade_flags).flatten()
        total_points_processed += len(lat_flat)
        # --- Create the scientific filter mask for L4A ---
        mask_valid = np.isfinite(agb_flat) & (agb_flat > -9999) # Valid agbd value
        mask_l4 = (l4_flag_flat == 1) # Good L4A Quality
        mask_degrade = (degrade_flag_flat == 0) # No signal degradation
        final_mask = mask_valid & mask_l4 & mask_degrade
        # Apply the mask
        lat_filtered = lat_flat[final_mask]
        lon_filtered = lon_flat[final_mask]
        agb_filtered = agb_flat[final_mask]
```

```
    if len(agb_filtered) == 0:
    tqdm.write(f"ℹ️ No high-quality points in {file_name}. Skipping.")
    continue
    # --- Create GeoDataFrame from filtered data ---
    gdf = gpd.GeoDataFrame(
    {"agbd": agb_filtered, "year": year, "month": month}, # Target is AGBD
    geometry=gpd.points_from_xy(lon_filtered, lat_filtered),
    crs="EPSG:4326"
    )
    # --- Clip to AOI ---
    gdf_clip = gpd.sjoin(gdf, aoi, how="inner", predicate="within")
    if len(gdf_clip) > 0:
    records.append(gdf_clip[["geometry", "agbd", "year", "month"]])
    total_points_kept += len(gdf_clip)
    tqdm.write(f"✅ {file_name} -> {len(gdf_clip)} high-quality points inside AOI.")
    except Exception as e:
    tqdm.write(f"❌ Error reading {file_name}: {e}")
    # === 5. MERGE ALL AND SAVE ===
    if not records:
    print("\n❌ No GEDI L4A data points found inside Yogyakarta.")
    else:
    print(f"\nMerging {len(records)} files...")
    final_gdf = pd.concat(records, ignore_index=True)
    final_gdf["lon"] = final_gdf.geometry.x
    final_gdf["lat"] = final_gdf.geometry.y
    final_df = final_gdf[["lon", "lat", "agbd", "year", "month"]] # Save agbd
    final_df.to_csv(output_csv, index=False)
    print("\n--- Summary ---")
    print(f"Total points processed: {total_points_processed}")
    print(f"Total high-quality points kept: {total_points_kept}")
    print(f"🎉 Done! Saved {len(final_df)} points to {output_csv}")
```

A.3 Merge All Dataset

```
    import pandas as pd
    import geopandas as gpd
    import rasterio
    import rasterio.sample
    import rasterio.features
    from shapely.geometry import box
    import os
    import numpy as np
    from tqdm import tqdm
    import glob # Import glob
    # === 1. DEFINE FILE PATHS ===
```

```
    BASE_DIR = os.path.dirname(os.path.abspath(__file__))
    # --- INPUTS ---
    # 1. Your 'ground truth' GEDI L4A (AGBD) points for 2020
    GEDI_CSV = os.path.join(BASE_DIR, "gedi_l4a_2020_filtered.csv")
    # 2. Your MASTER RASTER file pattern
    raster_file_pattern = os.path.join(BASE_DIR, "Yogyakarta Rasters 2020*.tif")
    # 3. Your Land Cover Map (for filtering)
    LANDCOVER_TIF                    =                    os.path.join(BASE_DIR,
"ESA_WorldCover_10m_Yogyakarta_2020.tif")
    if not os.path.exists(LANDCOVER_TIF):
    LANDCOVER_TIF_v2 = os.path.join(BASE_DIR, "WorldCover Yogyakarta 2020.tif")
    if os.path.exists(LANDCOVER_TIF_v2):
    LANDCOVER_TIF = LANDCOVER_TIF_v2
    else:
    print(f"❌ ERROR: Land Cover TIF not found at either path.")
    exit()
    # --- OUTPUT ---
    # 4. The final, most powerful dataset for your thesis
    OUTPUT_CSV = os.path.join(BASE_DIR, "FINAL_THESIS_DATASET_AGBD_2020.csv")
    # === 2. CHECK IF ALL FILES EXIST ===
    master_raster_files = glob.glob(raster_file_pattern)
    if not master_raster_files:
    print(f"❌ ERROR: No Master Raster files found.")
    print(f" Looking for file pattern: '{raster_file_pattern}'")
    print(" Make sure your downloaded TIF files are in this folder.")
    exit()
    print(f"✅ Found {len(master_raster_files)} master raster tiles.")
    if not os.path.exists(GEDI_CSV):
    print(f"❌ ERROR: GEDI CSV file not found: {GEDI_CSV}")
    exit()
    print(f"✅ Found Land Cover TIF: {os.path.basename(LANDCOVER_TIF)}")
    print("✅ All input files found.")
    # === 3. LOAD GEDI L4A POINTS (THE BASE DATASET) ===
    print(f"Loading GEDI L4A (AGBD) points from {GEDI_CSV}...")
    gedi_df = pd.read_csv(GEDI_CSV)
    master_gdf = gpd.GeoDataFrame(
    gedi_df,
    geometry=gpd.points_from_xy(gedi_df.lon, gedi_df.lat),
    crs="EPSG:4326" # GEDI points are in WGS84
    )
    print(f"Loaded {len(master_gdf)} total GEDI L4A points.")
    # === 4. SAMPLE MASTER RASTER (HANDLES MULTIPLE TILES) ===
    print("Sampling Master Raster (S1, S2, SRTM, ALOS)...")
    all_raster_samples = []
```

```
    band_names = []
    for tile_path in tqdm(master_raster_files, desc="Processing Raster Tiles"):
    with rasterio.open(tile_path) as src:
    tile_crs = src.crs
    tile_bounds_geom = gpd.GeoDataFrame(geometry=[box(*src.bounds)], crs=tile_crs)
    gdf_reprojected = master_gdf.to_crs(tile_crs)
    points_in_tile = gpd.sjoin(gdf_reprojected, tile_bounds_geom, how='inner',
predicate='within')
    if points_in_tile.empty:
    tqdm.write(f" No points found in tile {os.path.basename(tile_path)}. Skipping.")
    continue
    tqdm.write(f" Found {len(points_in_tile)} points in
{os.path.basename(tile_path)}.")
    coords = [(p.x, p.y) for p in points_in_tile.geometry]
    if not band_names:
    band_names = list(src.descriptions)
    if not band_names or all(b is None for b in band_names):
    # Fallback band names from your GEE script
    band_names = [
    'B2', 'B3', 'B4', 'B5', 'B6', 'B7', 'B8', 'B8A', 'B11', 'B12',
    'NDVI', 'EVI', 'SAVI', 'MSAVI2', 'NDMI', 'NBR', 'GNDVI', 'NDRE',
    'VV', 'VH', 'VH_asm', 'VH_contrast', 'VH_corr', 'VH_var',
    'VH_idm', 'VH_savg', 'VH_ent',
    'elevation', 'slope', 'aspect',
    'HH', 'HV', 'HV_HH_Ratio'
    ]
    print(f" Warning: No band names found. Using {len(band_names)} default names.")
    sampled_array = np.vstack(list(src.sample(coords)))
    sampled_df = pd.DataFrame(sampled_array, columns=band_names,
index=points_in_tile.index)
    all_raster_samples.append(sampled_df)
    if not all_raster_samples:
    print("❌ ERROR: No GEDI points were found inside *any* of the raster tiles.")
    exit()
    master_raster_df = pd.concat(all_raster_samples)
    # Replace infinities with NaN so they can be filled
    master_raster_df.replace([np.inf, -np.inf], np.nan, inplace=True)
    print("✅ Master Raster sampling complete.")
    # === 5. SAMPLE LAND COVER (SINGLE FILE) ===
    print(f"Sampling Land Cover map: {os.path.basename(LANDCOVER_TIF)}...")
    with rasterio.open(LANDCOVER_TIF) as src:
    gdf_reprojected = master_gdf.to_crs(src.crs)
    coords = [(p.x, p.y) for p in gdf_reprojected.geometry]
    try:
```

```
    lc_band_name = src.descriptions[0] or 'landcover'
    except (IndexError, TypeError):
    lc_band_name = 'landcover'
    if 'Map' in lc_band_name: lc_band_name = 'landcover' # Standardize
    sampled_array = np.vstack(list(src.sample(coords)))
    lc_df = pd.DataFrame(sampled_array, columns=[lc_band_name],
index=master_gdf.index)
    print("✅ Land Cover sampling complete.")
    # === 6. MERGE ALL DATASETS ===
    print("Merging all datasets...")
    # 'how="inner"' automatically drops points that are not in ALL datasets
    final_dataset = master_gdf.join(master_raster_df, how="inner").join(lc_df,
how="inner")
    # === 7. --- DEBUG STEP --- ===
    print("\n--- DEBUG: Land Cover Value Counts (Pre-filtering) ---")
    print("ESA WorldCover Legend:")
    print(" 10 = Trees, 20 = Shrubland, 30 = Grassland, 40 = Cropland")
    print(" 50 = Built-up, 60 = Barren, 80 = Open Water, 0 = NoData")
    print("----------------------------------------")
    if 'landcover' in final_dataset.columns:
    print(final_dataset['landcover'].value_counts())
    else:
    print("❌ CRITICAL ERROR: 'landcover' column not found in final_dataset!")
    print("----------------------------------------\n")
    # ==================================
    # === 8. FILTER BY VEGETATED CLASSES AND SAVE ===
    VEGETATED_CLASSES = [10, 20, 30, 40, 100] # Trees, Shrub, Grassland, Farm,
Mangrove
    print(f"Filtering for ALL vegetated classes: {VEGETATED_CLASSES}")
    # Also filter out any 'NoData' (0) values from the landcover
    vegetated_df = final_dataset[
    final_dataset['landcover'].isin(VEGETATED_CLASSES) &
    (final_dataset['landcover'] != 0)
    ].copy()
    # Drop the 'landcover' column as it's no longer needed
    vegetated_df = vegetated_df.drop(columns=['landcover'])
    if vegetated_df.empty:
    print("\n❌ ERROR: No GEDI points were found in the specified vegetated areas.")
    print(" Check the 'DEBUG' table above. If you only see 50, 60, or 80,")
    print(" then all your GEDI points are on 'Built-up', 'Barren', or 'Water'.")
    exit()
    # --- THIS IS THE FIX ---
    # Instead of dropping rows with *any* NaN, we will
    # fill the few missing pixel values (NaNs) with 0.
```

```
    # A Random Forest can easily handle 0s.
    original_count = len(vegetated_df)
    vegetated_df = vegetated_df.fillna(0)
    final_count = len(vegetated_df)
    print(f" Filled NaNs in {original_count - final_count} rows (if any). Kept
{final_count} points.")
    # ---------------------
    if vegetated_df.empty:
    print("\n❌ ERROR: All vegetated points were dropped.")
    exit()
    # Save the final file
    vegetated_df.to_csv(OUTPUT_CSV, index=False)
    print("\n--- Summary ---")
    print(f"Total original L4A points: {len(gedi_df)}")
    print(f"Points found in raster tiles: {len(master_raster_df)}")
    print(f"Total points kept (after filtering): {len(vegetated_df)}")
    print(f"🎉 Success! Final master dataset saved to: {OUTPUT_CSV}")
    print("\n➡️ Next step: Re-run your final training script
(model_training_RFE_SVR_MLP.py) using this new 2020 CSV file!")
```

A.4 Model Training Script

```
import pandas as pd
import numpy as np
import os
import joblib
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.neural_network import MLPRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import RFE, SelectKBest, mutual_info_regression,
f_regression
from sklearn.decomposition import PCA
from sklearn.metrics import r2_score, mean_squared_error
import matplotlib.pyplot as plt

# ============================================================
#   SETTINGS
# ============================================================
FEATURE_SELECTION_METHOD = "RFE"
# Options: "RFE", "MI", "KBest", "PCA"

N_SELECTED_FEATURES = 15
TEST_SIZE = 0.2
RANDOM_STATE = 42

# ============================================================
#   FILE PATHS
# ============================================================
BASE_DIR = os.getcwd()
DATASET_CSV = os.path.join(BASE_DIR, "FINAL_THESIS_DATASET_AGBD_2020.csv")
SCALER_FILE = os.path.join(BASE_DIR, "final_scaler_agbd_2020.joblib")
FS_FILE = os.path.join(BASE_DIR, "final_feature_selector_agbd_2020.joblib")
```

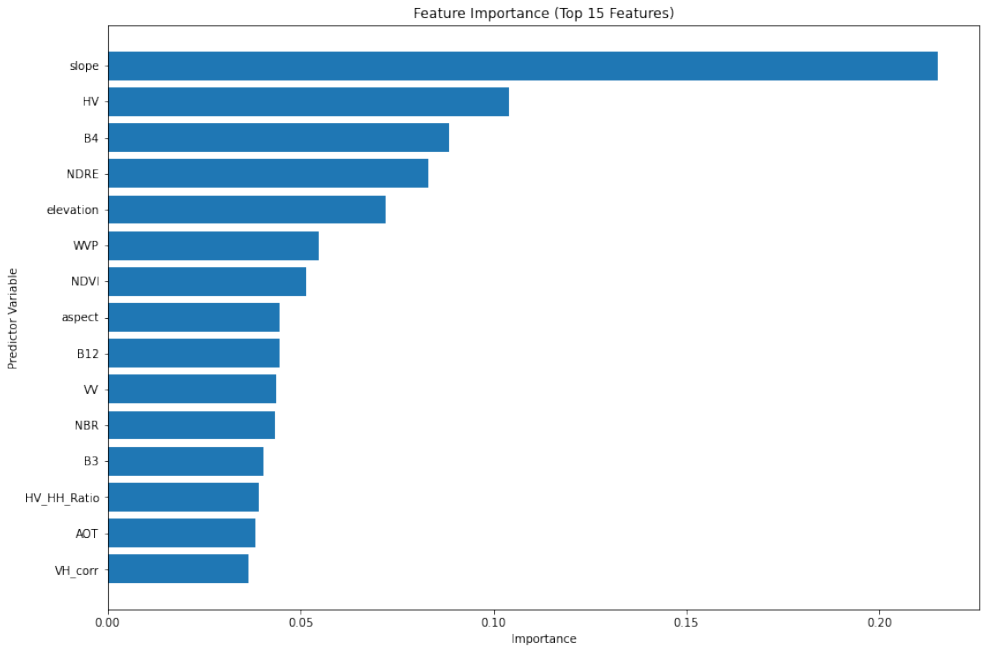
```
# ============================================================
#   LOAD DATA
# ============================================================
print(f"Loading dataset: {DATASET_CSV}")
data = pd.read_csv(DATASET_CSV).replace([np.inf, -np.inf], np.nan).dropna()

gedi_info_columns = ["lon", "lat", "agbd", "year", "month", "geometry"]
X = data.drop(columns=gedi_info_columns, errors="ignore")
y = data["agbd"]

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=TEST_SIZE, random_state=RANDOM_STATE
)

# ============================================================
#   SCALING
# ============================================================
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
original_feature_names = X.columns.tolist()

# ============================================================
#   FEATURE SELECTION METHODS
# ============================================================
print(f"\n=== Applying Feature Selection: {FEATURE_SELECTION_METHOD} ===")

if FEATURE_SELECTION_METHOD == "RFE":
    base_model = RandomForestRegressor(n_estimators=80, random_state=42)
    selector = RFE(base_model, n_features_to_select=N_SELECTED_FEATURES)
    X_train_fs = selector.fit_transform(X_train_scaled, y_train)
    X_test_fs = selector.transform(X_test_scaled)
    selected_idx = selector.get_support(indices=True)
    selected_features = [original_feature_names[i] for i in selected_idx]

elif FEATURE_SELECTION_METHOD == "MI":
    selector = SelectKBest(mutual_info_regression, k=N_SELECTED_FEATURES)
    X_train_fs = selector.fit_transform(X_train_scaled, y_train)
    X_test_fs = selector.transform(X_test_scaled)
    selected_idx = selector.get_support(indices=True)
    selected_features = [original_feature_names[i] for i in selected_idx]

elif FEATURE_SELECTION_METHOD == "KBest":
    selector = SelectKBest(f_regression, k=N_SELECTED_FEATURES)
    X_train_fs = selector.fit_transform(X_train_scaled, y_train)
    X_test_fs = selector.transform(X_test_scaled)
    selected_idx = selector.get_support(indices=True)
    selected_features = [original_feature_names[i] for i in selected_idx]

elif FEATURE_SELECTION_METHOD == "PCA":
    selector = PCA(n_components=N_SELECTED_FEATURES, random_state=42)
    X_train_fs = selector.fit_transform(X_train_scaled)
    X_test_fs = selector.transform(X_test_scaled)
    selected_features = [f"PCA_{i+1}" for i in range(N_SELECTED_FEATURES)]

else:
    raise ValueError("Invalid FEATURE_SELECTION_METHOD")

print("\nSelected Features:")
print(selected_features)

joblib.dump(selector, FS_FILE)
joblib.dump(scaler, SCALER_FILE)

# ============================================================
```

```
#   MODEL DEFINITIONS
# ============================================================
mlr = LinearRegression()
rf = RandomForestRegressor(n_estimators=120, random_state=42)
svr = SVR(kernel="rbf", C=10, epsilon=0.2)
mlp = MLPRegressor(hidden_layer_sizes=(120, 80),
                   max_iter=600,
                   random_state=42,
                   early_stopping=True)

models = {
    "MLR": mlr,
    "RF": rf,
    "SVR": svr,
    "MLP": mlp
}

results = {}

# ============================================================
#   TRAINING LOOP
# ============================================================
print("\n=== Training Models ===")
for name, model in models.items():
    print(f"\nTraining {name}...")
    model.fit(X_train_fs, y_train)
    y_pred = model.predict(X_test_fs)
    y_pred = np.clip(y_pred, 0, None)

    r2 = r2_score(y_test, y_pred)
    rmse = np.sqrt(mean_squared_error(y_test, y_pred))

    results[name] = {"R2": r2, "RMSE": rmse, "Model": model}

    print(f"   R² = {r2:.4f}")
    print(f"   RMSE = {rmse:.4f}")


# ============================================================
#   SELECT BEST MODEL
# ============================================================
best_model_name = max(results, key=lambda m: results[m]["R2"])
best_model = results[best_model_name]["Model"]

joblib.dump(best_model, os.path.join(BASE_DIR,
f"best_agbd_2020_{best_model_name}.joblib"))
print(f"\n🏆 Best Model: {best_model_name} (R² =
{results[best_model_name]['R2']:.4f})")

# ============================================================
#   PLOT: BEST MODEL SCATTER
# ============================================================
y_pred_best = best_model.predict(X_test_fs)
y_pred_best = np.clip(y_pred_best, 0, None)

plt.figure(figsize=(9, 7))
plt.scatter(y_test, y_pred_best, alpha=0.5)
plt.plot([y_test.min(), y_test.max()],
         [y_test.min(), y_test.max()], 'r--')
plt.xlabel("Observed AGBD")
plt.ylabel("Predicted AGBD")
plt.title(f"Best Model: {best_model_name} with {FEATURE_SELECTION_METHOD}\nR² =
{results[best_model_name]['R2']:.4f}")
plt.grid()
plt.savefig("best_model_scatter.png")
plt.close()
```

```
print("\n🎉 Completed All Training & Feature Selection!")
```

Appendix B: Model Outputs and Supplementary Visuals

B.1 Feature Importance List



Appendix C: Dataset Samples

C.1 GEDI Data Sample

lon	lat	agbd	year	month
110.12905901482841	-7.747665456684594	307.23224	2020	7
110.12935278886043	-7.7480802928888925	353.0187	2020	7
110.12965704734398	-7.748499123224214	149.2235	2020	7
110.129970270793009	-7.748921756083449	53.991352	2020	7
110.130263249123565	-7.74933623723512	86.228516	2020	7
110.130574429419461	-7.749757901224952	200.10052	2020	7
110.1308862298919	-7.750179803506982	111.935196	2020	7
110.13119543761964	-7.750600668965381	177.50145	2020	7
110.13149644692011	-7.751018285761764	142.89517	2020	7
110.13210875358955	-7.7518575563705564	195.17949	2020	7
110.13273025162829	-7.752700453033841	188.15054	2020	7
110.133073522546414	-7.753137477012977	225.50903	2020	7
110.133380370007534	-7.753554640287556	160.14026	2020	7
110.13368694284198	-7.753974423624018	259.47424	2020	7
110.13398617318617	-7.754391302720938	178.04926	2020	7
110.13429014871862	-7.754810059798437	143.26193	2020	7
110.13459284749402	-7.755228308147466	148.41289	2020	7
110.13489522356676	-7.755646433606982	158.78915	2020	7
110.135204886358046	-7.7560674341229605	230.83914	2020	7
110.13552064349925	-7.756490847270469	218.61208	2020	7

C.2 Master Dataset Sample

lon	lat	agbd	year	month	NDVI	EVI	VV	VH_entropy	elevation	slope	aspect	…
110.12906	-7.7476655	307.23224	2020	7	0.7952	0.3675	-10.6959	200.1542	723.47	10.49	38.13	…
110.12935	-7.7480803	353.01870	2020	7	0.7989	0.4274	-11.7903	192.0625	728.09	14.35	40.31	…
110.12966	-7.7484991	149.22350	2020	7	0.6259	0.3210	-13.8513	123.6250	726.20	11.90	99.01	…
110.12997	-7.7489218	53.99135	2020	7	0.8766	0.7031	-10.2949	165.5000	717.76	6.51	169.51	…
110.13026	-7.7493362	86.22852	2020	7	0.8785	0.6366	-7.3749	219.8958	712.61	5.42	211.53	…