MULTIMODAL DEEP LEARNING ARCHITECTURE FOR BIOMASS PREDICTION USING AERIAL IMAGERY

by

Sadikul Alim Toki

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Electrical Engineering

Approved:

Calvin Coopmans, Ph.D. Major Professor	Mario Harper, Ph.D. Committee Member
Sierra Young, Ph.D. Committee Member	Alfonso Torres-Rua, Ph.D. Committee Member

David F. Feldon, Ph.D.
Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2026

Copyright © Sadikul Alim Toki 2026

All Rights Reserved

ABSTRACT

MULTIMODAL DEEP LEARNING ARCHITECTURE FOR BIOMASS PREDICTION USING AERIAL IMAGERY

by

Sadikul Alim Toki, Master of Science

Utah State University, 2026

Major Professor: Calvin Coopmans, Ph.D.
Department: Electrical and Computer Engineering

In the field of precision agriculture, estimation of biomass (like the Leaf Area Index – LAI) is a key issue. The main challenge for this is to collect biomass data from the ground through human labor, and collecting measurements per plant is resource-consuming in terms of time and money. It is still manageable if it is over a small area like UAS coverage, as they have smaller altitudes, resulting in a plant-level resolution. However, it can be difficult to collect data for an area of coverage like that of satellite imagery, as the spatial coverage for those images can be very high while the detail information (pixel size) can be quite coarse and represent multiple plants or conditions. This study meets this challenge by training a convolutional neural network (CNN) model on a set of multispectral UAS imagery and showing its correlation with Landsat imagery. Each image has its respective ground-truth biomass measurements, i.e., Leaf Area Index (LAI) values, from which the model will learn about the spatial patterns of the images taken from grapevines across California as part of the USDA-led GRAPEX project and the AggieAir UAS Program. As each image has four radiometric spectral bands (R, G, B, and NIR) and corresponding pixel-wise height data i.e., Digital Surface Model (DSM) and Canopy Height Model (CHM), the CNN models will be trained on different combinations of spectral bands and height data. The models will be used

to generate continuous LAI maps, which have a pixel resolution of 3.6 meters (grapevine row spacing distance). Then the UAS-generated LAI maps are resampled to a 30-meter resolution map to match the format of the Landsat imagery resolution, to understand the association of model-predicted LAI values with Landsat bands and some other vegetation indices. Through the outcome of this research, it will be clear which combination of bands works best for predicting biomass from grape vineyards based on different ranges of the presence of biomass. Also, this research contributes to bridging the gap between UAS and satellite-based remote sensing, as this work provides insights into how UAS imagery-based LAI can provide biomass prediction from field/sensors scale to the grapevine plant scale and to the plot/vineyard scale, and thereby, for larger areas compatible with satellite imagery.

(84 pages)

PUBLIC ABSTRACT

MULTIMODAL DEEP LEARNING ARCHITECTURE FOR BIOMASS PREDICTION USING AERIAL IMAGERY

Sadikul Alim Toki

Grape growers rely on knowing how much leaf area their vines have, since canopy size affects how much water a vine needs, how healthy it is, and how much fruit it will produce. Measuring this by hand, plant by plant, across an entire vineyard takes enormous time and labor, so researchers have turned to cameras mounted on drones and satellites as a faster alternative. Drones capture very detailed images but only over small areas per flight, while satellites can photograph an entire farming region at once but cannot make out individual plants.

This thesis builds a computer vision model, a program trained to interpret images the way a person would recognize objects in a photo, that automatically estimates leaf area from drone photographs of two California vineyards, using ordinary color images together with invisible near-infrared light and camera-based measurements of plant height. Two different ways of combining these image types were built and compared to find the more accurate approach. The resulting model was used to create complete leaf-area maps across whole vineyard fields, which were then compared against satellite images of the same fields to check how well the two agreed. The goal is a practical tool that could one day help growers monitor vineyard health across large areas without walking every row and measuring every vine by hand.

To my father and mother,
everything I have built rests on what you built first

And to my brother, exactly as he is.

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Dr. Calvin Coopmans, for his guidance throughout my time at AggieAir. He gave me the freedom to pursue my own questions and the discipline to see them through, and his careful reading of countless drafts improved both this thesis and my research writing in ways I am still discovering.

I thank the members of my committee, Dr. Alfonso Torres-Rua, Dr. Sierra Young, and Dr. Mario Harper, for their time, their questions, and the suggestions that reshaped the direction of this work in numerous ways.

I am grateful to Richard Snider for helping me solve uncounted technical problems over the course of this research, and to Sergio Alvarez Carrion for sharing his dataset and guiding me through the GIS work I was not yet proficient in. My thanks also to Htoo Poo, Michael, and Emma for the shared troubleshooting, the second opinions, and the everyday conversations that made this a good place to work.

To my Bangladeshi friends in Logan: you turned a cold valley into something like home. Thank you for standing by me through the good stretches and the bad ones, and for the endless plates of Bangladeshi delicacies that arrived exactly when I needed them. Logan was a long way from home until you made it otherwise.

Finally, I am dedicating this thesis to my parents and my brother. No page of acknowledgement could add much to that. Thank you for everything that came before it.

Sadikul Alim Toki

CONTENTS

Page

ABSTRACT . . . iii

PUBLIC ABSTRACT . . . v

ACKNOWLEDGMENTS . . . vii

LIST OF TABLES . . . x

LIST OF FIGURES . . . xi

ACRONYMS . . . xiv

1 INTRODUCTION . . . 1

2 RELATED WORK . . . 4
2.1 Leaf Area Index: Importance and Measurement Methods . . . 4
2.2 Deep Learning in Remote Sensing . . . 5
2.3 Machine Learning and Deep Learning Approaches for LAI Estimation . . . 6
2.4 Biomass and Yield Prediction Using Leaf Area Index . . . 7
2.5 Multimodal Deep Learning and Fusion Strategies . . . 8
2.6 Summary and Research Gap . . . 10

3 DATASET DESCRIPTION . . . 12
3.1 UAS Platform and Imaging Sensor . . . 13
3.2 LAI Data Collection . . . 14
3.3 DSM and CHM computations . . . 15
3.4 Image Pre-Processing . . . 17
3.5 Experimentation Setup . . . 18
3.6 Architecture of the model . . . 20
3.7 Landsat Data Acquisition . . . 24

4 RESULTS AND DISCUSSION . . . 26
4.1 Performance on the Test Set . . . 26
4.2 Training and Validation Curves Analysis . . . 28
4.3 Test Performance Analysis on BAR and RIP Datasets . . . 39

5 INFERENCE ON A COMPLETE UAS FLIGHT DATA AND SATELLITE DATA 46
5.1 Field-Scale LAI Mapping on a whole UAS flight . . . 46
5.1.1 Field-Scale Inference Pipeline . . . 46
5.1.2 Predicted LAI Maps Across Channel Combinations . . . 49
5.1.3 Quantitative Summary of Predicted LAI Across the Flight . . . 53
5.2 Cross-Scale Validation of UAS-Derived LAI Against Landsat 8 Surface Reflectance and Vegetation Indices . . . 56

6 CONCLUSION 61
6.1 Conclusions 61
6.2 Future Work 63

REFERENCES 66

LIST OF TABLES

Table Page

3.1 Description of the dataset 18

3.2 Total trainable parameters per channel combination and fusion strategy . . 24

4.1 Early vs. Late Fusion test-set performance for images with $LAI \geq 3$ ($n = 158$). Within each combination's row, the better-performing fusion strategy for a given metric is bolded. 26

4.2 Early vs. Late Fusion test-set performance for images with $LAI < 3$ ($n = 352$). Within each combination's row, the better-performing fusion strategy for a given metric is bolded. 27

LIST OF FIGURES

Figure Page

1.1 End-to-end workflow of this thesis: UAS and field data collection, channel co-registration and preprocessing, parallel training of Early and Late Fusion Multi-Channel CNNs across eight channel combinations, held-out evaluation, field-scale LAI mapping, and satellite-scale cross-validation against Landsat 8. 3

3.1 The first image from the left shows the geolocation of the site where the dataset was collected from. The middle image is of Barelli (BAR), and the right one is of Ripperdan (RIP). The red markers indicate the location of the chips 13

3.2 GreatBlue sUAS platform 13

3.3 Architecture of the Early Fusion Multi-Channel CNN Model 22

3.4 Architecture of the Late Fusion Multi-Channel CNN Model 23

4.1 Training and Validation Curves for Early RGB combination 29

4.2 Training and Validation Curves for Early RGB-NIR combination 30

4.3 Training and Validation Curves for Early RGB-DSM combination 30

4.4 Training and Validation Curves for Early RGB-CHM combination 31

4.5 Training and Validation Curves for Early RGB-NIR-DSM combination . . . 31

4.6 Training and Validation Curves for Early RGB-NIR-CHM combination . . . 32

4.7 Training and Validation Curves for Early RGB-DSM-CHM combination . . 32

4.8 Training and Validation Curves for Early All Channels combination 33

4.9 Training and Validation Curves for Late RGB combination 33

4.10 Training and Validation Curves for Late RGB-NIR combination 34

4.11 Training and Validation Curves for Late RGB-DSM combination 34

4.12 Training and Validation Curves for Late RGB-CHM combination 35

4.13 Training and Validation Curves for Late RGB-NIR-DSM combination . . . 35

4.14 Training and Validation Curves for Late RGB-NIR-CHM combination . . . 36

4.15 Training and Validation Curves for Late RGB-DSM-CHM combination . . . 36

4.16 Training and Validation Curves for Late All Channels combination 37

4.17 Predicted vs. actual LAI on the two held-out test flights, BAR (left) and RIP (right), pooled across all eight channel combinations, for Early and Late Fusion 40

4.18 Scatter Plot for Test Results on BAR and RIP images by RGB-only combination model 40

4.19 Scatter Plot for Test Results on BAR and RIP images by RGB-NIR combination model 41

4.20 Scatter Plot for Test Results on BAR and RIP images by RGB-DSM combination model 41

4.21 Scatter Plot for Test Results on BAR and RIP images by RGB-CHM combination model 42

4.22 Scatter Plot for Test Results on BAR and RIP images by RGB-NIR-DSM combination model 42

4.23 Scatter Plot for Test Results on BAR and RIP images by RGB-NIR-CHM combination model 43

4.24 Scatter Plot for Test Results on BAR and RIP images by RGB-DSM-CHM combination model 43

4.25 Scatter Plot for Test Results on BAR and RIP images by All Channels combination model 44

5.1 Field-scale inference pipeline: an orthomosaic is built per channel combination from the target UAS flight, cut into the same 25×25 px (15 cm pixel^{-1}, 3.6 m $\times$ 3.6 m footprint) chip grid used in training, and passed through each trained model to produce one scalar LAI value per chip. Each value is written back out as a uniformly-filled single-band chip, and the resulting chip collection is assembled into a VRT mosaic and materialized into a single field-scale GeoTIFF, independently for all 16 models. 47

5.2 Study site, Madera County, California 49

5.3 Predicted LAI under **Early Fusion** for eight channel combinations 50

5.4 Predicted LAI under **Late Fusion** for eight channel combinations 51

5.5 Predicted LAI (mean ± std across all 15,162 chips) for each of the 16 trained models applied to the full Ripperdan flight, by channel combination and fusion strategy. 54

5.6 Relationship between UAS-derived LAI and Landsat 8 surface reflectance for the RGB channel-combination model 56

5.7 UAS-derived LAI versus Landsat 8 vegetation indices: NDVI and EVI for the RGB model 57

5.8 Spatial distribution at 30 m of UAS-derived LAI from the early-fusion and late-fusion models RGB channel-combination model, alongside the co-registered Landsat 8 NDVI and EVI 58

5.9 Relationship between UAS-derived LAI and Landsat 8 surface reflectance for the RGB-NIR channel-combination model 58

5.10 UAS-derived LAI versus Landsat 8 vegetation indices: NDVI and EVI for the RGB-NIR model 59

5.11 Spatial distribution at 30 m of UAS-derived LAI from the early-fusion and late-fusion models RGB-NIR channel-combination model, alongside the co-registered Landsat 8 NDVI and EVI 60

ACRONYMS

BAR	Barelli (GRAPEX vineyard site)
CHM	Canopy Height Model
CNN	Convolutional Neural Network
CRR	canopy relief ratio
DSM	Digital Surface Model
DTM	Digital Terrain Model
EPSG	European Petroleum Survey Group (geodetic coordinate reference registry)
ET	evapotranspiration
EVI	Enhanced Vegetation Index
FWHM	full width at half maximum
GDAL	Geospatial Data Abstraction Library
GIS	Geographic Information System
GRAPEX	Grape Remote Sensing Atmospheric Profiling and Evapotranspiration
GSD	ground sample distance
LAI	Leaf Area Index
LaSRC	Land Surface Reflectance Code
LiDAR	Light Detection and Ranging
MAE	Mean Absolute Error

ML	Machine Learning
MSE	Mean Squared Error
NDVI	Normalized Difference Vegetation Index
NIR	Near-Infrared
PCA	Plant Canopy Analyzer
ReLU	Rectified Linear Unit
RIP	Ripperdan (GRAPEX vineyard site)
RMSE	Root Mean Squared Error
RS	Remote Sensing
SAR	Synthetic Aperture Radar
sUAS	small Uncrewed Aerial System
TAI	Tree Area Index
UAS	Uncrewed Aerial System
USGS	United States Geological Survey
UTM	Universal Transverse Mercator
VI	Vegetation Index
VRT	Virtual Raster (GDAL mosaic format)
VTOL	Vertical Takeoff and Landing
WGS	World Geodetic System

CHAPTER 1

INTRODUCTION

Over the past few decades, remote sensing has become an essential tool, helping people make decisions in everything from environmental monitoring and water management to land use and agriculture [1], [2]. In the early days, these analyses depended on getting satellite data down to ground stations for further study [3].

The process has gotten much faster thanks to new remote sensing technology and advances in machine learning [4]. Still, sending all that data to the ground isn't always easy: there are limited communication windows, bad weather can mess with the signal, and sometimes technical issues get in the way [5], [6]. Because of these challenges, there's been a shift toward doing more of the data crunching right on the satellite itself. That way, only the most important results get sent down, which saves a lot of bandwidth [7], [8].

For nearly twenty years, researchers at Utah State University's AggieAir program have been at the forefront of this work, coming up with practical solutions for water management and precision farming. Their research covers everything from designing small drones for efficient data collection to developing software that helps process the scientific data those drones collect [9].

A big focus in precision agriculture is figuring out how much plant material, or biomass, is in a field. This is often measured by a quantity called the Leaf Area Index (LAI), which is the total area of leaves divided by the ground area underneath [10], [11]. Getting accurate LAI numbers takes a lot of work: traditional field measurements mean either picking apart plants by hand or carefully measuring them, which is slow, labor-intensive, and hard to repeat throughout the season [12]. Remote sensing provides a better way, but each method has its own pros and cons. Small drones can fly low and capture details down to individual plants or vine rows, giving results almost as good as ground surveys, while covering much more ground in one go. On the other hand, satellites can scan an entire region in a single pass

but their images are so coarse that each pixel can cover several plants or different conditions, making them less useful for measuring biomass at the plant or plot level [13], [14].

The work in this thesis aims to bridge that gap by producing biomass estimates with the fine detail of drone imagery, but in a way that can scale up to the bigger area covered by satellites. The study predicts LAI using six different types of data collected during each drone flight: Red, Green, Blue, Near-Infrared (NIR), Digital Surface Model (DSM), and Canopy Height Model (CHM). The AggieAir lab has brought together its strengths in drone design and data processing to build end-to-end systems for predicting biomass using drone imagery. They are currently working on a multi-channel deep learning model to estimate the biomass in grape vineyards based on LAI, which basically measures how much greenery exists in a given area [10]. The model works best with certain combinations of these data types, depending on the LAI range.

This thesis builds on that work by comparing two strategies for combining the six types of input data. One method, called Early Fusion, stacks all the channels together so the model processes them at once. The other, called Late Fusion, has the model handle each group of channels (like RGB, NIR, DSM, and CHM) separately before combining the results. Both strategies are tested across all possible data combinations. The trained models aren't just judged on test samples: they're also used to make continuous LAI maps for whole fields. These maps are then scaled up to match the 30-meter resolution of Landsat 8 satellite images and compared against satellite vegetation data, addressing the challenge of connecting drone-scale and satellite-scale predictions. Figure 1.1 shows the full workflow, from collecting data in the field and with drones, through all the processing steps, model training, mapping, and finally, cross-checking with satellite data.

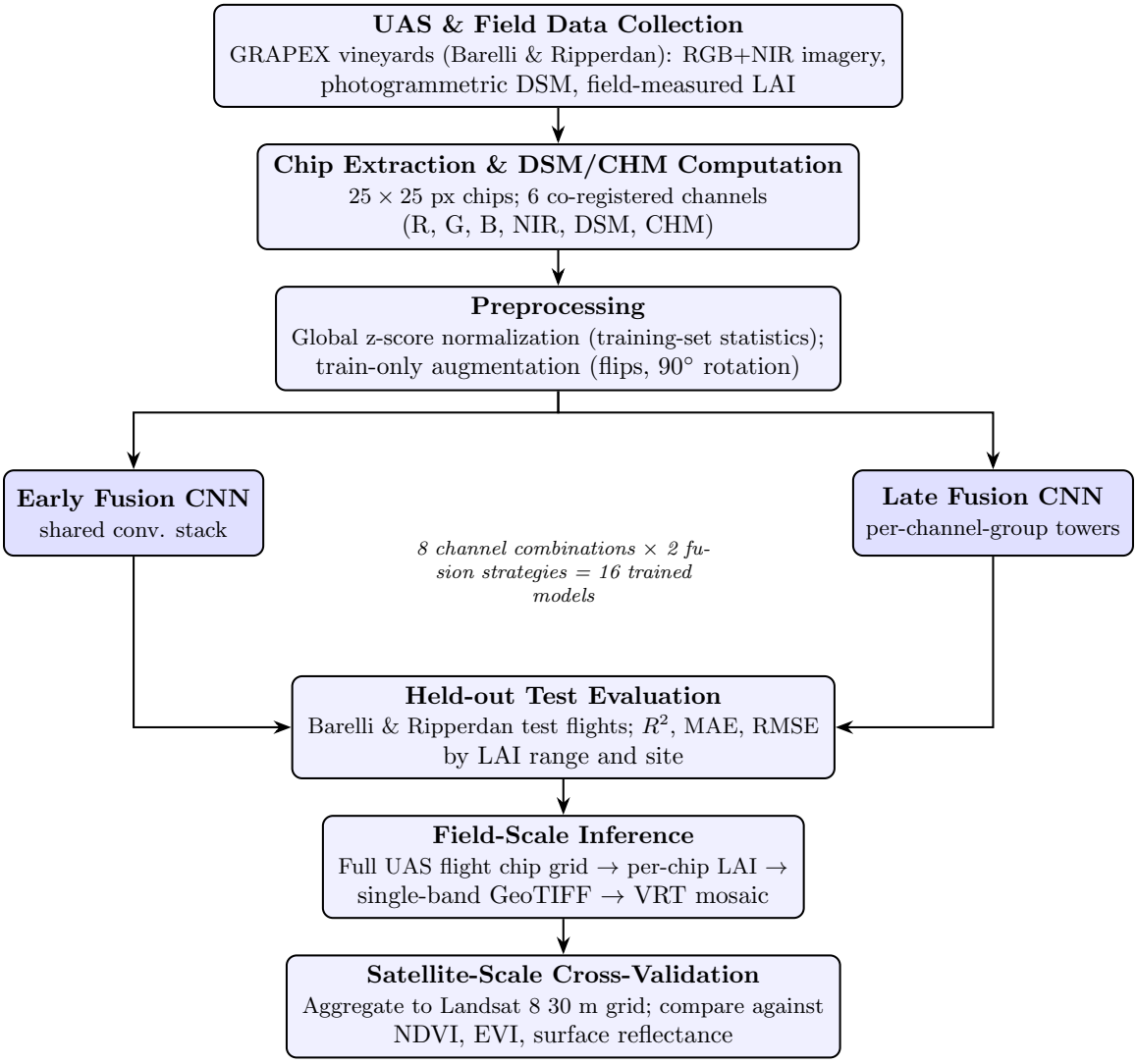


Fig. 1.1: End-to-end workflow of this thesis: UAS and field data collection, channel co-registration and preprocessing, parallel training of Early and Late Fusion Multi-Channel CNNs across eight channel combinations, held-out evaluation, field-scale LAI mapping, and satellite-scale cross-validation against Landsat 8.

CHAPTER 2

RELATED WORK

2.1 Leaf Area Index: Importance and Measurement Methods

Leaf Area Index (LAI) is a key biophysical parameter in describing vegetation structure, since it directly governs the canopy's radiative transfer, that is, how much of the incident solar radiation is intercepted, absorbed, and scattered back toward a sensor rather than reaching the soil background, which is precisely the physical basis that makes LAI retrievable from the spectral reflectance measured by optical remote sensing platforms [11]. In particular, LAI is a key parameter in evapotranspiration (ET) partitioning [15], and consequently in irrigation scheduling and precision agriculture more broadly, which is what motivates the sustained research interest in measuring it accurately and at scale.

Remote Sensing (RS) has become an alternative to ground-based LAI measurement through the use of passive optical sensors, and active sensors such as microwave and Light Detection and Ranging (LiDAR) [16]. The RS approach allowed the development of new methodologies to compute LAI, such as the use of vegetation indices (VI) and Machine Learning (ML). Ground-based indirect measurement remains important as well, both as a data source and as the calibration and validation reference against which any remote-sensing-based LAI product must be checked. White et al. [17] evaluate several indirect optical protocols using a Plant Canopy Analyzer against destructively sampled, direct LAI measurements in California vineyards, finding that a protocol facing the canopy with readings spread evenly across the interrow space gave the strongest agreement with direct measurements. However, even the best-performing indirect protocol still requires visiting each plot in person, which is exactly the coverage bottleneck remote sensing is positioned to address: not to replace or outperform ground-based measurement, but to extend it, using indirect optical measurement of the kind evaluated by White et al. as the calibration

and validation reference against which remote-sensing-based LAI is checked, while scaling estimation to whole fields or vineyards that could never be measured this densely by hand.

2.2 Deep Learning in Remote Sensing

Automated decision-making systems are now significantly more efficient due to the rapid development of remote sensing technologies, facilitated by deep learning [4]. Zhu et al. [18] provide one of the foundational comprehensive reviews of this shift, cataloging how convolutional architectures came to outperform traditional feature-engineering pipelines across remote sensing tasks such as scene classification, object detection, and semantic segmentation, while also highlighting challenges specific to remote sensing data, including sensor heterogeneity, limited labeled training data, and the scale mismatch between satellite, aerial, and ground observations. Osco et al. [19] narrow this review specifically to UAS-acquired imagery, the platform most relevant to this thesis, and summarize how deep learning classification and regression techniques have been applied to UAS data across agricultural, environmental, and urban monitoring applications, while noting that UAS-specific constraints, such as limited training samples per flight and variable flight and illumination conditions, still require careful handling that differs from satellite-scale deep learning.

While convolutional architectures have dominated the approaches surveyed above, a substantial and more recent line of work has instead adapted the transformer architecture, originally developed for natural language processing, to remote sensing imagery. Bazi et al. [20] apply a Vision Transformer directly to remote sensing scene classification, showing that self-attention over image patches can match or exceed CNN-based baselines without any convolutional layers at all. Hong et al. [21] and Wang et al. [22] extend this transformer-based direction to hyperspectral image classification specifically, with the latter combining a convolutional feature extractor with a transformer-based context module rather than replacing convolutions outright, illustrating that hybrid CNN-transformer designs are also an active area of the field. Most recently, this transformer-based direction has converged with self-supervised pretraining to produce large remote sensing foundation models, trained on hundreds of millions of unlabeled images and then fine-tuned for a specific

downstream task [23], a paradigm shift away from the task-specific, fully supervised training used throughout this thesis and the studies reviewed in this chapter, but one that still depends on exactly the kind of well-curated, task-specific labeled data this thesis contributes for fine-tuning and evaluation.

Within the specific sub-task of semantic segmentation, most relevant to the land cover classification work by the AggieAir group discussed below, Li et al. [24] review over 160 deep learning segmentation methods for remote sensing imagery and, notably, identify precision agriculture as one of the primary application domains driving continued methodological development in this area. Deep learning has also been directly compared against, and combined with, transfer learning for related land cover and land use classification tasks; Dastour and Hassan [25] compare several deep transfer learning methods for land use and land cover classification, finding that models pretrained on large natural-image datasets and fine-tuned on a smaller labeled remote sensing dataset can reach competitive accuracy without training a large architecture from scratch, an approach that is particularly attractive when labeled remote sensing data, as in this thesis, is comparatively scarce.

Within this broader trend, the AggieAir group at Utah State University has directly contributed to applying deep learning to remote sensing tasks relevant to this thesis. Toki et al. [26] compare three deep learning semantic segmentation architectures, UNet++, Feature Pyramid Network, and Pyramid Scene Parsing Network, for land cover classification, and evaluate their robustness under various image deformations and augmentations, a question of practical importance for any deep learning pipeline, including the one in this thesis, that is expected to generalize beyond its own training imagery to a full, previously unseen UAS flight.

2.3 Machine Learning and Deep Learning Approaches for LAI Estimation

A substantial body of recent research has applied machine learning, and to a lesser extent deep learning, directly to the problem of estimating LAI itself from remotely sensed or ground-based sensor data. A time-series-based approach named TMEAD-BiLSTM predicts the LAI of alfalfa plants, incorporating auxiliary factors like temperature and soil moisture

alongside the time series itself, distinguishing it from classical, purely spectral regression models [27].

Several studies instead rely on 3D structural information derived from LiDAR. One study measures LAI from UAS-based LiDAR data, deriving structural parameters such as height percentiles and the canopy relief ratio (CRR), and trains five different machine learning algorithms on these features, finding that more complex ML algorithms achieved significantly higher accuracy [28]. LiDAR-derived structural features have also been used to train a multi-scale LAI estimation model based on multiple linear regression, motivated by a shortage of fine-scale, ground-truthed LAI data [14]. Apart from UAS-based approaches, ground-based LiDAR scanning has also been used to estimate LAI directly; this research suggests that 40–50 scans are needed to calculate the Tree Area Index (TAI) with adequate precision [29].

Other studies rely primarily on multispectral or RGB imagery rather than structural (LiDAR) information. A group of researchers combined imagery of several plant species, Alfalfa, Tall Fescue, and Plantain, together with multiple vegetation indices, to train several ML models, finding that Random Forest had superior prediction capability among the models compared [30]. In another study, regression models trained on eight different vegetation indices showed that the red and NIR bands were the most effective for predicting peanut LAI [31]. A separate study used only RGB imagery to predict LAI in rice, comparing a Plant Canopy Analyzer (PCA)-based approach against regression models built on color indices and on vegetation indices [32].

2.4 Biomass and Yield Prediction Using Leaf Area Index

A closely related, but distinct, line of research uses LAI not as the final quantity of interest but as an intermediate variable from which crop biomass or yield is then estimated. Dong et al. [13] estimate LAI for six crop types from harmonized Landsat 8 and Sentinel-2 imagery using vegetation indices, and then use the resulting LAI time series to reparametrize a simple crop growth model, showing that remotely sensed LAI can serve as an effective intermediate variable for regional-scale crop biomass estimation rather than

requiring biomass to be predicted directly from raw spectral bands. Not every biomass-estimation study routes through LAI in this way, however; Togeiro de Alckmin et al. [33] instead retrieve simulated hyperspectral features directly from multispectral UAS data and use them, together with conventional vegetation indices, as inputs to a random forest model to estimate perennial ryegrass pasture biomass, finding comparable accuracy between the two feature sets without an explicit LAI estimation step.

Within viticulture specifically, canopy vigor and structural predictors closely related to LAI have been used to model site-specific vineyard outcomes. Taylor et al. [34] collect three years of yield, pruning mass (vine size), canopy vigor, and soil data across 321 sites in a single vineyard, and compare stepwise linear regression against random forest regression for predicting site-specific yield and pruning mass, finding that the more complex, non-linear relationship between canopy vigor and yield was better captured by random forest, while the simpler relationship for pruning mass was adequately captured by linear regression. This distinction, that a canopy structural or vigor proxy can predict some vineyard outcomes more easily than others, and that the appropriate model complexity can differ by outcome, echoes the pattern found later in this thesis for LAI-based prediction across different channel combinations and LAI ranges.

2.5 Multimodal Deep Learning and Fusion Strategies

Instead of just using one type of sensor, a lot of research has explored multimodal deep learning. These models combine information from different sources, such as various sensors, data types, or spectral ranges, so everything is analyzed together rather than separately. Baltrušaitis et al. [35] formalize the core challenges shared by this line of work, representation, translation, alignment, fusion, and co-learning, and note that how and where the information from different modalities is combined, the fusion strategy, is one of the most consequential architectural decisions in a multimodal model, since it determines whether cross-modal correlations can be learned jointly or are only combined after each modality has already been processed independently. Li et al. [36] adapt this taxonomy specifically to remote sensing, organizing deep learning-based multimodal fusion into three levels: pixel- or

input-level fusion, in which raw bands from different sources are stacked into a single input volume before any learned feature extraction; feature-level fusion, in which each modality is processed by its own sub-network and the resulting feature representations are combined partway through the model; and decision-level fusion, in which each modality is processed by an entirely separate model and only the final outputs are combined. In the multi-channel CNN literature, input-level fusion is commonly referred to as *early fusion*, while feature- or decision-level fusion, where each modality keeps its own dedicated processing pathway for most or all of the network, is referred to as *late fusion*.

Both strategies have concrete precedent in remote sensing. Zhou et al. [37] systematically review fusion strategies and deep learning architectures specifically for UAS-acquired multimodal remote sensing data, RGB, LiDAR, SAR, and hyperspectral among them, across applications including precision agriculture, and find that no single fusion strategy or architecture is reported as consistently superior across studies, reinforcing that the choice remains an empirical, task-specific question rather than a settled one. Within that broader picture, Jiang et al. [38] propose a double-channel CNN for multispectral RGB-NIR image classification, motivated by the observation that the red, green, and blue bands are strongly correlated with one another but only weakly correlated with the near-infrared band; rather than stacking all four bands into one input, they process RGB and NIR through two separate convolutional networks and fuse the resulting features only at the fully connected layer, a late-fusion design intended to let each sub-network specialize in the statistics of its own bands. By contrast, Irfan et al. [39] directly compare early and late fusion for a land use and land cover classification task combining Synthetic Aperture Radar (SAR) and optical imagery, and find that stacking both modalities into a single input volume and processing them with one shared network (early fusion) achieved higher overall accuracy (88%, F1-score 87%) than processing each modality with an independent network and combining the resulting decisions (late fusion, 84% accuracy, F1-score 82%) for that particular task. Taken together, these results indicate that neither fusion strategy is uniformly superior, and that the better choice depends on the specific modalities, task, and amount of training data

available, which motivates evaluating both strategies directly rather than assuming one in advance.

Multimodal fusion has also been applied directly to UAS-based crop and canopy trait prediction. Maimaitijiang et al. [40] fuse UAS-acquired RGB, multispectral, and thermal imagery, together with derived canopy structural and textural features, using input-, feature-, and decision-level fusion strategies to predict soybean grain yield, and find that a deep neural network combining modalities at the feature level outperformed models trained on any single modality alone. Nguyen et al. [41] similarly fuse UAS-acquired hyperspectral, thermal, and LiDAR data within a multi-task convolutional network to jointly predict several maize phenotyping traits, again finding that fusing multiple sensors improved on single-sensor baselines. These studies show that both the choice to fuse multiple UAS-derived data sources and the specific point in the network at which that fusion happens can materially affect prediction accuracy for vegetation traits, which is the same design question this thesis investigates for vineyard LAI: rather than committing to a single fusion strategy, Chapter 3 (Section 3.6) trains and directly compares an early-fusion and a late-fusion Multi-Channel CNN across every channel combination considered.

2.6 Summary and Research Gap

Across the areas reviewed above, LAI is well established as a biophysical parameter worth measuring accurately (Section 2.1), a wide range of ML techniques have been applied to estimating it from optical, multispectral, and LiDAR data (Section 2.3), and separately, LAI or closely related canopy vigor proxies have been used as an intermediate step toward predicting biomass and yield (Section 2.4). At the same time, deep learning has become the dominant approach for remote sensing tasks in general (Section 2.2) and multimodal fusion strategies have been directly compared for other remote sensing and agricultural tasks (Section 2.5). However, it can be seen that most of the research reviewed on LAI and biomass prediction specifically used classical machine learning models rather than deep learning approaches, and none of it directly compares early versus late fusion for this task. This thesis takes this gap as its motivation, building a multi-channel deep learning architecture

specifically for vineyard LAI prediction, directly comparing early and late fusion across every channel combination considered, and showing its implementation at field and satellite scale.

CHAPTER 3

DATASET DESCRIPTION

This chapter describes the UAS imagery dataset used to train and evaluate the LAI prediction models presented in this thesis, the UAS platform and imaging sensor used to collect it (Section 3.1), how the ground-truth LAI labels used to train and validate those models were collected in the field (Section 3.2), the computation of the Digital Surface Model (DSM) and Canopy Height Model (CHM) elevation models used as model inputs (Section 3.3), the pre-processing applied to every channel before training (Section 3.4), the experimental setup used to train the models (Section 3.5), the architecture of the two Multi-Channel CNN models compared in this work (Section 3.6), and, finally, how the Landsat 8 imagery used for the satellite-scale cross-validation in Chapter 5 was separately acquired (Section 3.7).

The dataset used in this study is a UAS Imagery dataset from the Grape Remote Sensing Atmospheric Profiling and Evapotranspiration (GRAPEX) project, comprising data from two vineyard sites in California's central valley, Barelli (BAR) and Ripperdan (RIP) (Figure 3.1). The dataset consists of images from 3 UAS flights over Barelli and 7 UAS flights over Ripperdan, taken at different times; the date and time of each flight are given in Table 3.1. A consistent naming convention was followed for each flight. For example, in "BAR_190627_1207", "BAR" is the abbreviation for the site the data was collected at, "190627" represents the collection date in "Year-Month-Date" format, and "1207" represents the collection time in 24-hour format. A total of 2293 chips were extracted from these 10 flights, where each chip has a separate LAI value. Each chip has 6 channels (Red, Green, Blue, Near Infrared (NIR), DSM, and CHM) in it, covers a 3.6 m x 3.6 m footprint, and has a pixel resolution of 15 cm x 15 cm.

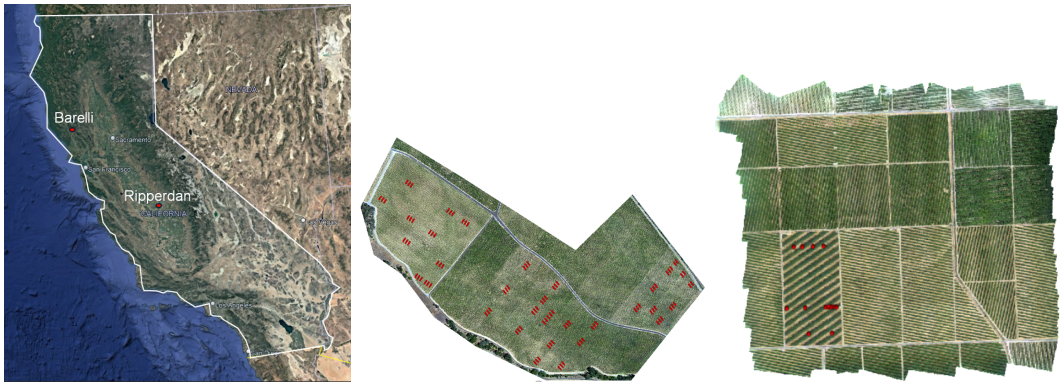


Fig. 3.1: The first image from the left shows the geolocation of the site where the dataset was collected from. The middle image is of Barelli (BAR), and the right one is of Ripperdan (RIP). The red markers indicate the location of the chips

3.1 UAS Platform and Imaging Sensor

All UAS imagery used in this study was collected using AggieAir's GreatBlue platform (Figure 3.2), an all-electric, 55-pound vertical-takeoff-and-landing (VTOL) transition fixed-wing small uncrewed aircraft system (sUAS) developed at Utah State University [42]. GreatBlue combines four vertically-oriented motors and propellers, used for VTOL launch and recovery, with a separate horizontally-oriented pusher motor and fixed wing used for efficient cruise flight, a quad-plane configuration that removes the need for a runway while retaining the endurance of a fixed-wing aircraft; the platform can remain airborne for up to 2.5 hours per flight while collecting multispectral remotely-sensed imagery.



Fig. 3.2: The AggieAir GreatBlue sUAS: on the ground in the AggieAir lab, showing the fixed wing and the four VTOL rotor booms (left), and in cruise flight (right). Reproduced from Coopmans et al. [42].

The imaging payload itself changed across the life of the AggieAir GRAPEX campaign, from an early consumer photography camera (a Canon S95, used in 2014) to a series of industrial narrowband cameras (Lumenera) in subsequent years. Every UAS flight used in this thesis was collected between 2018 and 2019 (Table 3.1), within the period, 2017 through 2021, during which the same Lumenera industrial narrowband sensor was used across every GRAPEX flight. This sensor captured four co-registered bands: Red (center wavelength 639.0 nm, full width at half maximum (FWHM) 42 nm), Green (center 542.5 nm, FWHM 27 nm), Blue (center 464.5 nm, FWHM 23 nm), and Near-Infrared (center 865.0 nm, FWHM 30 nm). These four spectral bands, together with the photogrammetrically-derived DSM and CHM described in Section 3.3, make up the six co-registered input channels used throughout this thesis.

3.2 LAI Data Collection

The ground-truth LAI label attached to every chip was not measured directly on the imagery; it was collected in the field, independently of the UAS flights, as part of the GRAPEX project [17]. Both Barelli and Ripperdan, the two sites used in this study, were among the vineyard sites where this field protocol was applied, so the same collection methodology underlies every LAI label used to train and evaluate the models in this thesis.

In the field, two different ways of measuring were used. One was an indirect, non-destructive optical method that made it easy to take routine, repeated samples. The other was a direct, destructive method applied to just a few plots, mainly to check and fine-tune the results from the indirect approach. The indirect method used a Plant Canopy Analyzer (PCA, LAI-2200C, LI-COR), a handheld instrument that estimates LAI from how much light is attenuated as it passes through the canopy. The sensor uses a fisheye lens to project a hemispherical image onto five concentric detector rings at different zenith angles, and a filter that excludes radiation above 490 nm to reduce the contribution of scattered light. At each plot, the instrument was used to take one above-canopy reading, immediately followed by a set of below-canopy readings taken about 30 cm above the ground; actual LAI is estimated from the logarithm of the resulting canopy gap fraction, averaged across the

below-canopy readings. Readings were generally taken under clear sky and away from solar noon to minimize the effect of direct sun, and additional above-canopy sky readings were collected so that a scattering correction could be applied during post-processing.

The team tried out seven different ways of collecting measurements, changing how the sensor was aimed (either at the canopy or along the row) and where exactly they took readings under the vines. They checked each method against direct, destructive LAI measurements to see which lined up best. The approach that worked best was aiming the sensor at the canopy and spreading the readings evenly across the width of the space between the rows. If the sensor was pointed along the row, or if readings were only taken underneath the canopy, the LAI numbers ended up being too low. A protocol using only four canopy-facing readings across the interrow performed nearly as well as one using ten readings ($R^2 = 0.87$ versus $R^2 = 0.88$, both with RRMSE around 15–16%), and was therefore recommended, and used, as the practical protocol for routine indirect LAI collection across the GRAPEX vineyards.

The direct, destructive measurements used to validate the indirect protocol were collected immediately after the corresponding indirect readings, at the same 1 m section of vine row. Every leaf in that section was removed at the petiole; a representative subset of the removed leaves, roughly 25% by count, was measured on a leaf area meter to obtain leaf area, and all of the removed leaves were then oven-dried to obtain dry weight. The ratio of measured leaf area to dry weight in the subset was used to scale the total dry weight of the full sample up to a total leaf area, which was then divided by the ground area of the sampled row section (row spacing multiplied by row length) to obtain direct LAI. Because this destructive method physically removes the sampled vines' foliage, it could only be applied sparingly and was used to validate the indirect PCA protocol rather than as the primary source of LAI labels for this study.

3.3 DSM and CHM computations

The DSM for each flight was generated from the UAS point cloud, and the CHM for that flight was then computed in Agisoft Metashape v2.1.0 as the difference between the DSM

and a corresponding bare-ground Digital Terrain Model (DTM), i.e., CHM = DSM − DTM.

Rather than computing a single DTM per study site, a separate DTM was generated for each individual flight. This was because comparing DSMs from different flights over the same site revealed spatial (x, y, z) coordinate discrepancies of up to approximately 2 m, an offset large enough to distort canopy-height estimates given the relatively short height of the vine canopy. Each per-flight DTM was derived from the flight's own point cloud using Metashape's ground-point classification tool, whose parameters (maximum angle, maximum distance, and maximum terrain slope) were manually tuned per flight based on visual inspection of the resulting surface. The classified DTM was then resampled to a coarser pixel size (on the order of a few meters), since terrain elevation in the vineyards changes smoothly and gradually; this suppressed small artificial sinks and depressions introduced by the classification step and reduced file size. All DTM and DSM rasters used a nodata value of −32767 and were referenced to the WGS 84 / UTM Zone 10N (EPSG:32610) coordinate system.

Ground-point classification was not always reliable directly out of Agisoft. In one recurring failure mode, actual vine canopy was misclassified as ground, which was corrected by tightening the maximum-distance parameter of the classifier. In a small number of cases, the point cloud for a given flight could not produce an acceptable DTM at all (e.g., large sunken regions of clearly incorrect elevation). For two such Ripperdan flights, the DTM was instead substituted either from a nearby flight's point cloud, or, in one case, computed independently by masking bare-soil pixels with an NDVI threshold and interpolating the terrain surface between them using ordinary kriging; this kriging-based DTM produced a more realistic CHM than reusing a borrowed DTM from another flight. Similarly, for one Ripperdan flight where the point-cloud-derived DSM itself was unreliable, an existing DSM already computed by the AggieAir group was substituted in its place. To check the quality of the CHM results, each map was compared by eye to trellis heights measured in the field. A few flights had obviously flat CHMs, where the canopy height was near zero across the whole field, and these were marked as problems. Most flights, though, produced CHMs

with canopy heights that lined up well with the expected trellis height, so they were kept for this study.

3.4 Image Pre-Processing

Each of the six channels (Red, Green, Blue, NIR, DSM, and CHM) was extracted from the GeoTIFF chips and normalized independently using a global z-score normalization. Rather than rescaling each chip using its own minimum, maximum, or percentile values, every channel was normalized using a single population-level mean and standard deviation computed once from the entire training split. For a given channel b, the per-pixel value I_b was standardized as

$$Z_b = \frac{I_b - \mu_b}{\sigma_b} \tag{3.1}$$

where μ_b and σ_b are the mean and standard deviation of channel b computed across every valid pixel of every chip in the training split only; the validation and test splits were held out of this calculation to prevent data leakage. Global, training-set-derived statistics were used instead of a per-chip minimum-maximum or percentile stretch specifically because the latter would rescale every chip, regardless of its true brightness or height, to fill the same output range, destroying exactly the kind of chip-to-chip brightness and height differences that carry information about canopy density. The resulting z-score was clipped to ± 3 standard deviations, a threshold chosen because it retains roughly 99.7% of pixel values under an approximately Gaussian channel distribution while suppressing the disproportionate influence of the small number of extreme values introduced by sensor noise, specular reflections, or DSM point-cloud reconstruction artifacts, then linearly rescaled to the range $[0, 1]$:

$$I_{b,\text{norm}} = \frac{\text{clip}(Z_b, -3, 3) + 3}{6} \tag{3.2}$$

Pixels flagged as nodata (i.e., outside the valid sensor footprint) were assigned a value of 0.5, the midpoint of the normalized range, rather than 0, since 0 would otherwise encode

a pixel three standard deviations below the training mean, a false signal for missing data. Because μ_b and σ_b are fixed constants derived once from the training set rather than recomputed per chip, each chip retains its brightness or height scale relative to the training population. This preserves the physical contrast between a low-vigor or bare-ground chip and a dense-canopy chip, which a per-chip adaptive stretch would otherwise remove by rescaling every chip, regardless of its true brightness, to fill the same output range. The same cached μ_b, σ_b statistics computed during training were reused unchanged when applying the trained models to the full UAS orthomosaic at inference time, keeping training-time and inference-time normalization consistent. The images were already of size 25 x 25 pixels.

3.5 Experimentation Setup

Table 3.1 shows the list of 10 flights used in this study.

The first eight flights were pooled together and partitioned into training and validation subsets using a fixed 80:20 split, held constant across all training runs rather than reshuffled by site or date. The remaining two flights, one from Barelli and one from Ripperdan, were withheld entirely from training and validation and used only to evaluate the trained models on unseen data.

Within the training partition only, each chip was augmented after normalization by adding a horizontal flip, a vertical flip, and a 90° rotation of itself, quadrupling the number

Table 3.1: Description of the dataset

Name	Number of Chips	Purpose of Use
BAR_190730_1028	490	Train + Validation
BAR_191627_1041	340	Train + Validation
RIP_180619_1229	170	Train + Validation
RIP_180713_1040	170	Train + Validation
RIP_180805_1044	171	Train + Validation
RIP_180805_1233	171	Train + Validation
RIP_180806_1041	171	Train + Validation
RIP_190504_1025	100	Train + Validation
BAR_190627_1207	340	Test
RIP_180619_1120	170	Test

of training samples relative to the raw chip count; the validation and test partitions were left unaugmented. These transformations were picked because they don't change the actual meaning of the data. When you flip or rotate a chip by 90°, the total leaf area and its true LAI value stay the same. That means every new, flipped, or rotated version still counts as a good training example, just like the original. Photometric changes like tweaking brightness or contrast were avoided on purpose. These kinds of adjustments would interfere with the actual channel values that the global z-score normalization (described above in Section 3.4) depends on to maintain a meaningful brightness and height scale across all chips. Adding artificial changes during training could end up making the model think that random brightness differences are connected to LAI, when that's not really the case. Every channel combination used in this study, RGB only; RGB+NIR; RGB+DSM; RGB+CHM; RGB+NIR+DSM; RGB+NIR+CHM; RGB+DSM+CHM; and all six channels together, was trained independently under both fusion strategies compared in this work, giving 16 models in total, each with its own weights, training history, and held-out test evaluation.

Each model was trained for up to 100 epochs with a batch size of 24 using the Adam optimizer. Training used the early stopping method, monitoring validation loss with a patience of 15 epochs and, once triggered, restoring the model weights from the epoch with the lowest validation loss. The learning rate was reduced by a factor of 0.2 whenever validation loss failed to improve for 5 consecutive epochs, down to a minimum learning rate of 1×10^{-7}. Independently of the weights restored by early stopping, the best-performing weights observed during training (lowest validation loss) were checkpointed separately, and the model state at the end of training was also saved on its own for comparison.

Taken together, these choices were made specifically to control overfitting given the comparatively small, augmented training set used in this study: early stopping and the accompanying learning-rate reduction stop training before a model can continue reducing training loss at the expense of validation performance; the 4× training-only augmentation described above increases the effective diversity of the training set without requiring additional field sites or flights; and, at the architecture level (Section 3.6), dropout and

batch normalization are applied throughout the convolutional stacks and fusion head of both fusion strategies. Section 4.2 evaluates directly how effective this combination was at controlling overfitting, particularly for the higher-capacity Late Fusion architecture.

All models were trained to minimize a robust LAI loss rather than a standard mean squared error (MSE) loss, chosen to reduce the influence of noisy or outlying LAI field measurements on training. This loss is a Huber-style function: for a residual $r = y_{\text{true}} - y_{\text{pred}}$ and a threshold δ, the per-chip loss is defined piecewise as

$$L_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| < \delta \\ \delta\left(|r| - \frac{1}{2}\delta\right), & |r| \geq \delta \end{cases} \tag{3.3}$$

and the loss minimized during training is the mean of $L_\delta(r)$ over each batch. For residuals smaller than δ, the loss is quadratic and behaves identically to MSE, so small prediction errors are still penalized smoothly; for residuals larger than δ, the penalty grows only linearly with the size of the error rather than quadratically, so a small number of chips with an unusually large residual, for example due to LAI measurement noise or a plot-to-chip misalignment, contribute proportionally less to the gradient than they would under MSE, without being excluded from training altogether. All 16 models in this study used $\delta = 0.5$ LAI units.

After training, each model was evaluated once on its held-out test flight. Performance was reported using the robust LAI loss (Equation 3.3, with the same $\delta = 0.5$ used in training) alongside MSE, root mean squared error (RMSE), mean absolute error (MAE), R^2, Pearson correlation, and three custom accuracy metrics: the fraction of predictions within 10% of the true LAI value, within 20% of the true LAI value, and within 0.5 LAI units of the true value. These metrics were computed separately for chips with a measured LAI below 3.0 and chips with a measured LAI at or above 3.0, so that model performance could be compared across the two LAI ranges rather than only in aggregate.

3.6 Architecture of the model

For every channel combination described in Section 3.5, two competing Multi-Channel CNN architectures, Early Fusion (Figure 3.3) and Late Fusion (Figure 3.4), were built and trained independently. Both architectures accept the same set of active 25×25 pixel inputs for a given combination (a 3-channel RGB input, and, when active, single-channel NIR, DSM, and CHM inputs), and both end in an identical fusion head and output layer; the two architectures differ only in how the active channels are convolved before that shared fusion head.

In Early Fusion, all active inputs are stacked together along the channel axis into a single $25 \times 25 \times n$ volume, where n is the number of active channels for that combination, and this combined volume is passed through one shared convolutional stack. In Late Fusion, each active channel group (RGB as one group, and NIR, DSM, and CHM each as their own group) is instead passed through its own independent convolutional stack, so that, for example, the all-channel combination has four separate towers rather than one. Every convolutional stack, whether the single shared stack in Early Fusion or an individual tower in Late Fusion, uses the same depth and filter sizes: two 3×3 convolutional layers of 32 filters each with ReLU activation, each followed by Batch Normalization; a 2×2 Max Pooling layer; a third 3×3 convolutional layer of 48 filters with ReLU activation, followed by Batch Normalization; and a Flatten layer that turns the resulting feature map into a one-dimensional feature vector. Keeping the stack architecture identical between the two fusion strategies means that any difference in parameter count or performance comes from the fusion strategy itself, that is, whether the channels share one set of convolutional filters or each channel group learns its own, rather than from a difference in convolutional depth. Because Late Fusion gives every active channel group its own stack, it has substantially more parameters than Early Fusion for every combination with more than one active channel group, growing to almost four times as many total parameters as Early Fusion for the all-channel combination (Table 3.2), while for single-group inputs such as RGB only, the two architectures reduce to the same stack and therefore have an identical parameter count.

The single flattened feature vector produced by Early Fusion, or the concatenation of

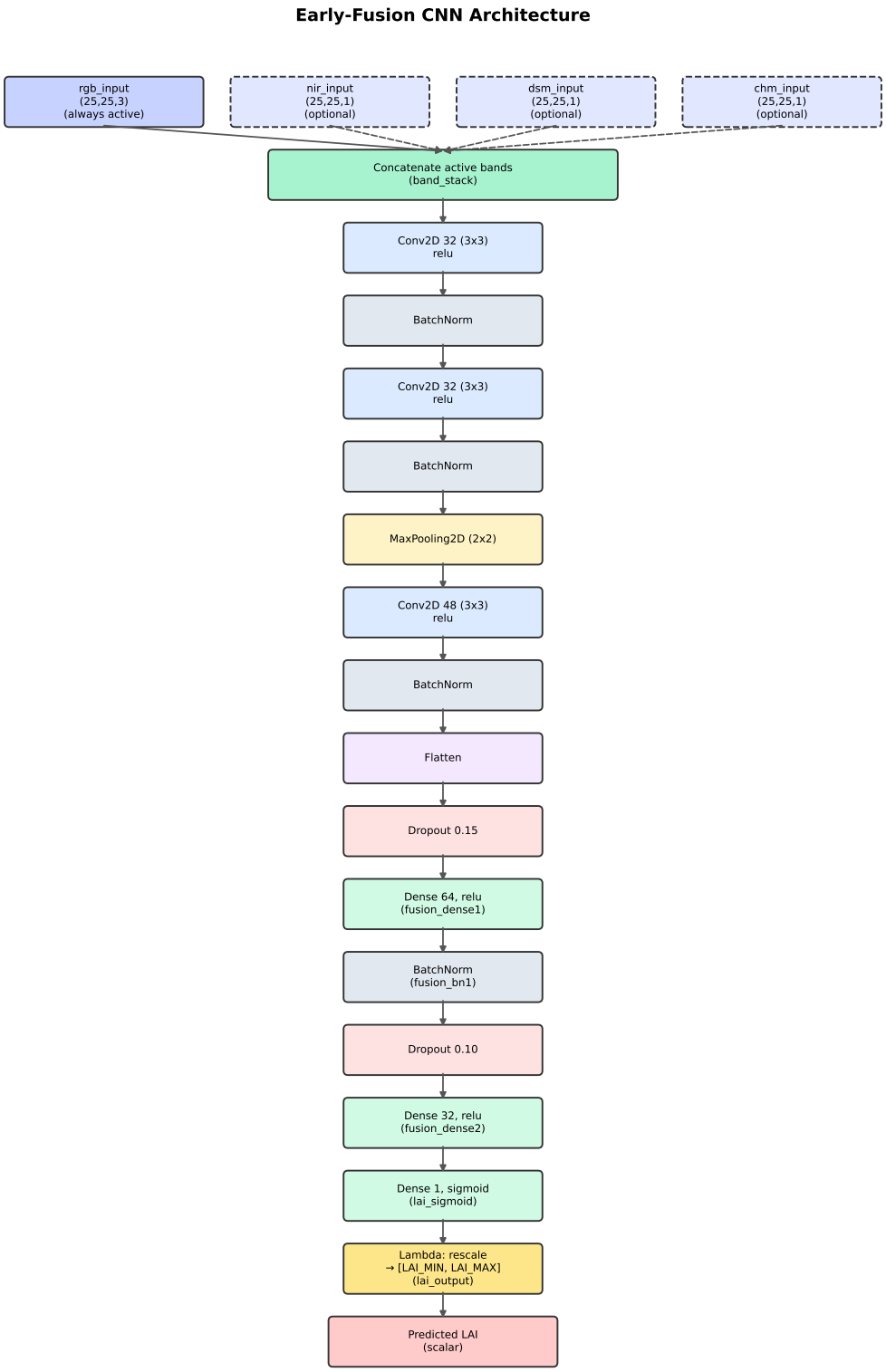


Fig. 3.3: Architecture of the Early Fusion Multi-Channel CNN Model

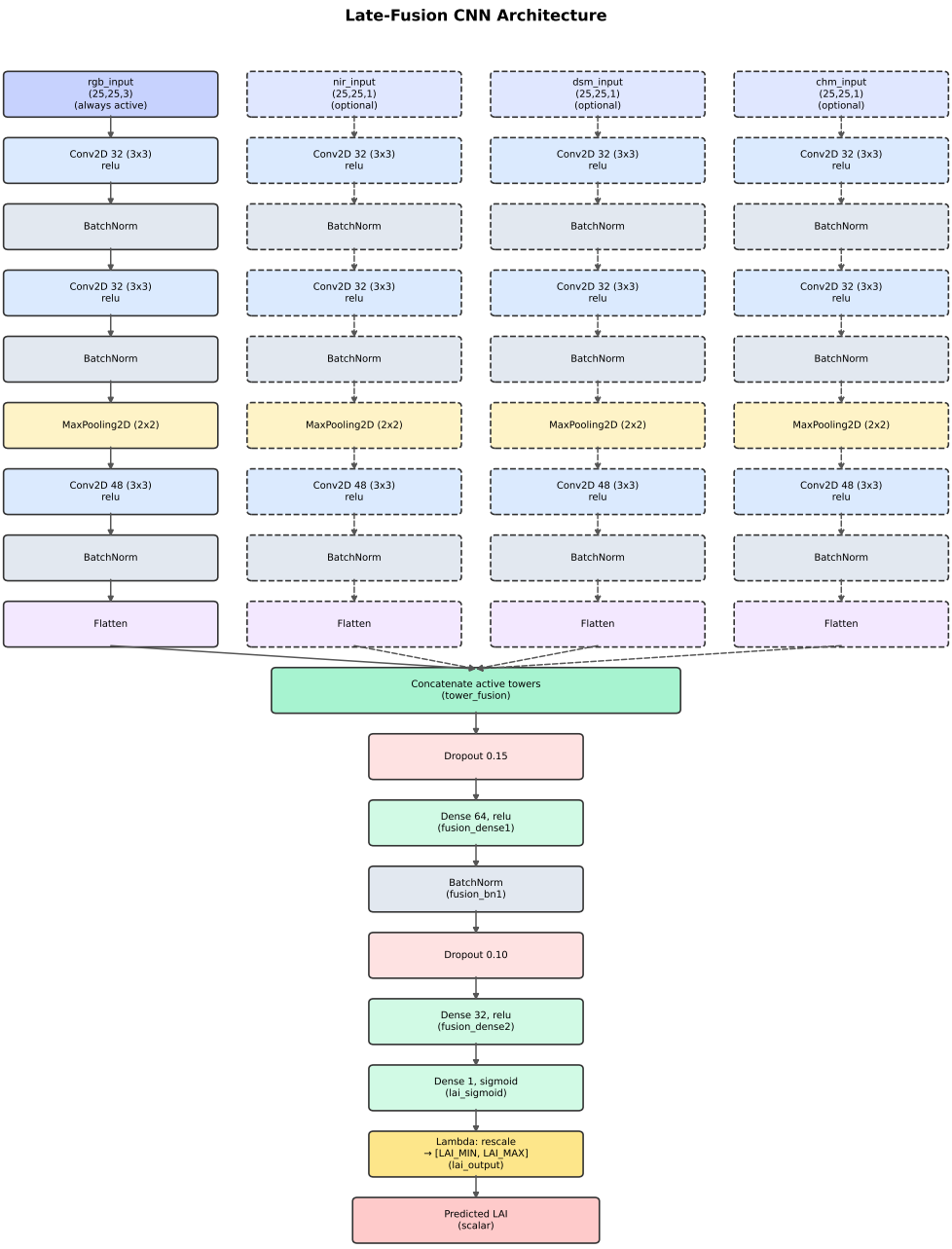


Fig. 3.4: Architecture of the Late Fusion Multi-Channel CNN Model

Table 3.2: Total trainable parameters per channel combination and fusion strategy

Combination	Channels	Early Fusion Params	Late Fusion Params
RGB	1 group	469,265	469,265
RGB-NIR	2 groups	469,553	935,521
RGB-CHM	2 groups	469,553	935,521
RGB-DSM	2 groups	469,553	935,521
RGB-NIR-DSM	3 groups	469,841	1,401,777
RGB-NIR-CHM	3 groups	469,841	1,401,777
RGB-DSM-CHM	3 groups	469,841	1,401,777
All Channels	4 groups	470,129	1,868,033

the per-group flattened feature vectors produced by Late Fusion, is then passed through an identical fusion head in both architectures: a Dropout layer that drops 15% of its inputs, a Dense layer of 64 units with ReLU activation, a Batch Normalization layer, a second Dropout layer that drops 10% of its inputs, and a final Dense layer of 32 units with ReLU activation. The output layer is a single Dense unit with a sigmoid activation, which is then rescaled from its native $[0, 1]$ range to the physically valid LAI range $[0, 5]$ using a fixed linear transformation. Bounding the output layer to this range this way ensures that the model can never predict an LAI value outside what is physically meaningful, which matters when the trained models are later applied to whole-orthomosaic chips that may fall outside the training distribution (e.g., bare ground or roads). Both architectures were compiled and trained identically, using the same optimizer, robust LAI loss, and evaluation metrics described in Section 3.5.

3.7 Landsat Data Acquisition

In addition to the UAS imagery described above, this thesis makes use of Landsat 8 imagery for the satellite-scale cross-validation presented in Chapter 5. Landsat scenes were obtained from the United States Geological Survey (USGS) EarthExplorer portal rather than searched by scene identifier or path/row directly. Instead, each search area of interest was defined using the bounding coordinates of the corresponding UAS flight footprint (Figure 3.1), so that the resulting Landsat 8 scene was guaranteed to fully cover the same Barelli or Ripperdan vineyard block already described in Section 3.2. For each UAS flight used in

the cross-scale comparison, the available Landsat 8 scene with an acquisition date closest to that flight's date was selected, minimizing the time elapsed between the two acquisitions and, with it, the risk that canopy condition or soil moisture changed meaningfully between the UAS overflight and the satellite overpass; scenes with substantial cloud cover over the study area were excluded from consideration.

Landsat Collection 2 Level-2 Science Products were downloaded rather than the corresponding Level-1 product [43]. Level-1 data represent raw, top-of-atmosphere radiance that has not been corrected for atmospheric scattering or absorption, whereas Level-2 products are radiometrically corrected to surface reflectance using the Land Surface Reflectance Code (LaSRC) algorithm together with auxiliary atmospheric data, so that the resulting reflectance approximates what would be measured at ground level in the absence of the intervening atmosphere. Because the UAS-derived LAI predictions compared against Landsat in Section 5.2 are themselves ground-referenced quantities, comparing them against uncorrected, top-of-atmosphere radiance would confound genuine surface-level agreement with atmospheric differences between the two acquisition dates; using the radiometrically corrected Level-2 product removes that confound and keeps the comparison consistent with how the vegetation indices used elsewhere in this thesis, NDVI and EVI, are conventionally computed from surface, rather than top-of-atmosphere, reflectance. From each selected scene, the Blue, Green, Red, and Near-Infrared surface reflectance bands were extracted at their native 30 m pixel resolution, both to compute NDVI and EVI and to serve as the direct spectral comparison bands used in Section 5.2.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Performance on the Test Set

All 16 trained models (Section 3.6) were evaluated once on the held-out test flights described in Section 3.5, BAR_190627_1207 and RIP_180619_1120, for a combined test set of 510 chips (158 with $LAI \geq 3$ and 352 with $LAI < 3$). Table 4.1 reports test performance for the high-LAI subset, and Table 4.2 reports the low-LAI subset, for every channel combination under both Early and Late Fusion.

Considered over the full test set without splitting by LAI threshold, Early Fusion RGB-CHM is the strongest model overall, with the highest R^2 (0.91), the lowest MSE (0.10) and MAE (0.24), and 87.1% of its predictions within ±0.5 LAI units. More broadly, Early Fusion models occupy 6 of the top 7 positions when all 16 models are ranked by full-test-set R^2; the only Late Fusion model to break into that group is RGB-NIR, in third place ($R^2 = 0.90$). Despite Late Fusion having up to almost four times as many parameters as Early Fusion for the same channel combination (Table 3.2), that extra per-modality capacity does not translate into better generalization to the two held-out test flights, suggesting Late Fusion is more prone to overfitting the comparatively small augmented training set (Section 3.5)

Table 4.1: Early vs. Late Fusion test-set performance for images with $LAI \geq 3$ ($n = 158$). Within each combination's row, the better-performing fusion strategy for a given metric is bolded.

Combination	MSE		MAE		Acc. within 10%		Acc. within 20%		Acc. within ±0.5	
	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late
RGB	**0.14**	0.14	**0.29**	0.29	**67.7**	66.5	**96.2**	95.6	81.6	**82.3**
RGB-NIR	0.15	**0.14**	0.31	**0.29**	66.5	**70.9**	93.7	**97.5**	82.9	**85.4**
RGB-CHM	**0.15**	0.16	**0.30**	0.31	63.3	**64.6**	93.7	**96.2**	**80.4**	76.6
RGB-DSM	0.18	**0.13**	0.35	**0.30**	57.6	**67.1**	88.6	**96.8**	72.2	**81.6**
RGB-NIR-DSM	**0.16**	0.21	**0.33**	0.38	**61.4**	53.2	**91.1**	86.7	**79.1**	69.6
RGB-NIR-CHM	**0.15**	0.18	**0.32**	0.33	**65.8**	62.0	**94.9**	92.4	**77.2**	75.9
RGB-DSM-CHM	**0.13**	0.15	**0.29**	0.31	**69.0**	63.9	**97.5**	94.9	**82.9**	77.2
All Channels	**0.15**	0.17	**0.32**	0.34	**60.8**	57.0	92.4	92.4	**79.7**	73.4

Table 4.2: Early vs. Late Fusion test-set performance for images with $LAI < 3$ ($n = 352$). Within each combination's row, the better-performing fusion strategy for a given metric is bolded.

Combination	MSE		MAE		Acc. within 10%		Acc. within 20%		Acc. within ±0.5	
	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late
RGB	**0.18**	0.20	**0.33**	0.35	**32.7**	31.8	53.7	**56.3**	**76.1**	72.4
RGB-NIR	0.11	**0.11**	0.25	**0.25**	**39.5**	38.9	68.8	**69.0**	89.2	89.2
RGB-CHM	**0.08**	0.11	**0.21**	0.25	**50.0**	39.2	**76.1**	71.6	**90.1**	87.8
RGB-DSM	0.27	**0.26**	0.42	**0.41**	**25.9**	25.0	46.0	**48.6**	64.5	**65.3**
RGB-NIR-DSM	**0.10**	0.15	**0.23**	0.29	**45.7**	34.9	**73.6**	64.5	**91.2**	84.4
RGB-NIR-CHM	**0.11**	0.12	**0.25**	0.27	**40.9**	36.1	**68.5**	67.6	**89.5**	85.8
RGB-DSM-CHM	**0.12**	0.15	**0.26**	0.30	**44.9**	37.8	**68.2**	60.8	**87.2**	81.2
All Channels	**0.10**	0.11	**0.23**	0.25	**44.9**	42.9	**73.3**	70.2	**92.0**	89.5

than Early Fusion is. The two weakest models overall are RGB-DSM under both fusion strategies ($R^2 = 0.78$ Early, $R^2 = 0.80$ Late); notably, both are worse than the RGB-only baseline in either fusion strategy, indicating that DSM alone did not just fail to help but actively hurt prediction quality relative to using RGB alone.

For the high-LAI subset ($LAI \geq 3$, Table 4.1), RGB-DSM-CHM under Early Fusion achieves the lowest MSE (0.13) and MAE (0.29), while RGB-NIR under Late Fusion achieves the highest accuracy on every threshold (70.9% within 10%, 97.5% within 20%, 85.4% within ±0.5 LAI), tying RGB-DSM-CHM Early for the best accuracy within 20%. No single channel combination or fusion strategy dominates every metric for dense-canopy chips, but canopy-structure information, whether from CHM or from NIR under Late Fusion, is consistently associated with the strongest performance.

For the low-LAI subset ($LAI < 3$, Table 4.2), the pattern is much less ambiguous. RGB-CHM under Early Fusion is the clear best performer, with the lowest MSE (0.08) and MAE (0.21) and the highest accuracy within both 10% (50.0%) and 20% (76.1%); the All Channels combination under Early Fusion only overtakes it nearly with 92.0% of predictions within ±0.5 LAI. RGB-DSM is again the weakest combination by a wide margin under both fusion strategies (MSE above 0.26, fewer than 26% of predictions within 10%), reinforcing that DSM alone is a poor predictor of LAI in low-vigor or sparsely-vegetated chips, where canopy height differences are least informative. Taken together, these results show that CHM is the most consistently valuable auxiliary channel across both LAI ranges,

DSM alone is consistently the least useful and can even be counterproductive, and Early Fusion generalizes to unseen flights at least as well as, and usually better than, Late Fusion despite its substantially smaller parameter count.

These results are best understood in terms of what DSM and CHM values physically represent. DSM records the absolute elevation of the first return above a fixed vertical datum at every pixel, whether that return comes from bare ground, a trellis wire, or the top of the canopy (Section 3.3); its values therefore mix two physically distinct quantities, the local terrain elevation, which reflects the site's own topography and has no causal relationship to LAI, and the height added by the canopy itself, which does. Because a separate per-flight DTM was required specifically to remove this terrain component, and cross-flight DSM discrepancies of up to 2 m were observed even at the same site (Section 3.3), the terrain component of DSM is neither small nor consistent from chip to chip or flight to flight; a model trained on raw DSM values must therefore learn to separate this arbitrary, flight-specific terrain signal from the genuine canopy-height signal buried within it, using only 2293 chips in total. CHM removes this confound before the model ever sees it: CHM = DSM − DTM isolates the canopy contribution directly, so every CHM value is already expressed relative to the local ground surface and is comparable across chips regardless of the site's underlying topography. This difference helps explain the results. CHM stands out as the most useful extra channel, while DSM on its own isn't just unhelpful, as it can actually hurt model performance, making RGB-DSM perform worse than using only RGB in both fusion strategies. Rather than a genuinely uninformative input, DSM is better understood as noisy and confounded, an interpretation that also anticipates the angled banding artifacts discussed in Section 5.1 that appear specifically in DSM-containing field-scale maps but not in CHM-only ones.

4.2 Training and Validation Curves Analysis

Every model's training run was monitored with the same six curves per epoch, training and validation robust LAI loss, MAE, and MSE (Section 3.5) in the top row, and accuracy within 10%, within 20%, and within ±0.5 LAI in the bottom row. Because early stopping

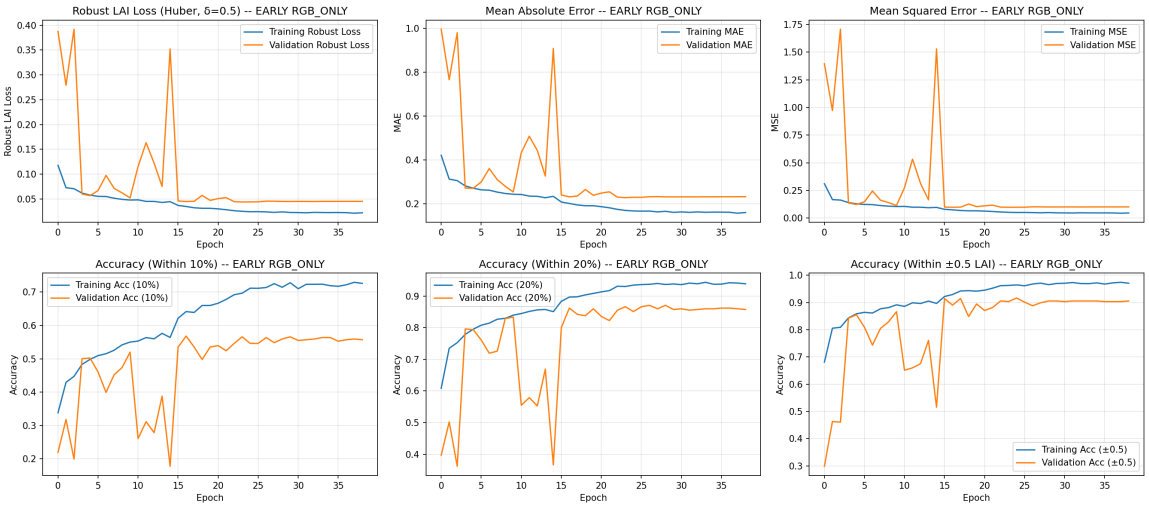


Fig. 4.1: Training and Validation Curves for Early RGB combination

halted each run independently once validation loss stopped improving for 15 consecutive epochs, the total number of epochs actually completed varies considerably across all 16 models, from as few as 29 epochs (RGB-DSM-CHM, Late Fusion) to more than 70 epochs (RGB-CHM, Late Fusion, Figure 4.12). This range is wide within both fusion strategies rather than being a simple early-versus-late split as the Early Fusion models ran from 37 to 59 epochs, while Late Fusion models ranged more broadly still, from 29 to 71 epochs.

A pattern common to every model, regardless of channel combination or fusion strategy, is a volatile opening phase followed by a much calmer remainder of training. For roughly the first 15–20 epochs, the validation curves show sharp, isolated spikes in robust loss, MAE, and MSE, together with corresponding drops in the accuracy panels, that are not mirrored in the training curves, which decrease comparatively smoothly from the first epoch (for example, RGB Early, Figure 4.1, and RGB-DSM-CHM Early, Figure 4.7, each show one or two such spikes before settling). Past this point, both training and validation curves flatten into a much smoother, near-monotonic trend for the remainder of training, and in no model configuration does validation loss trend back upward before training stops, indicating that the early stopping and learning-rate-reduction schedule (Section 3.5) was effective at halting training before validation performance began to degrade.

In every model, the training and validation loss lines (robust loss, MAE, MSE) usually

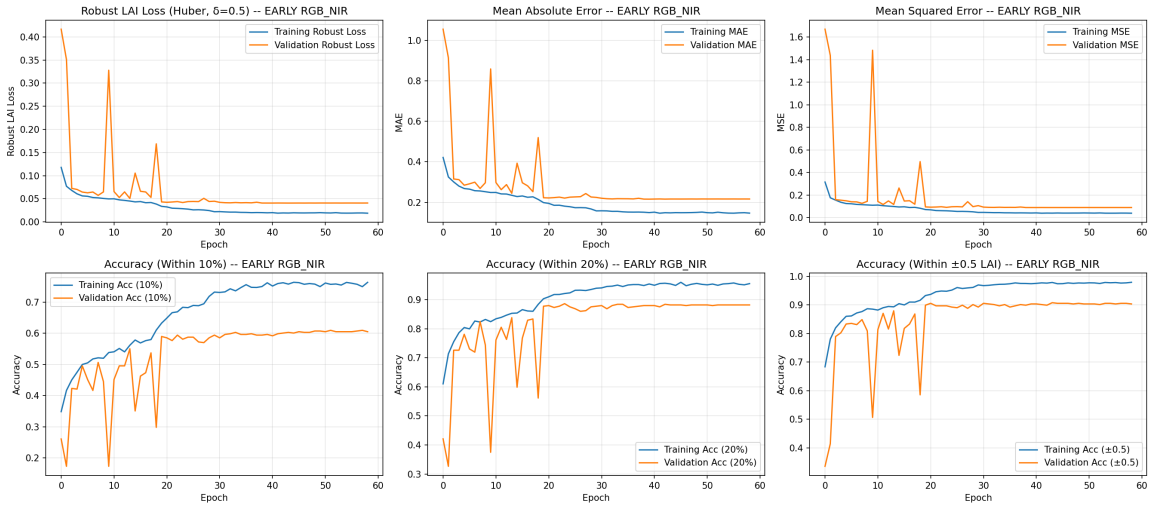


Fig. 4.2: Training and Validation Curves for Early RGB-NIR combination

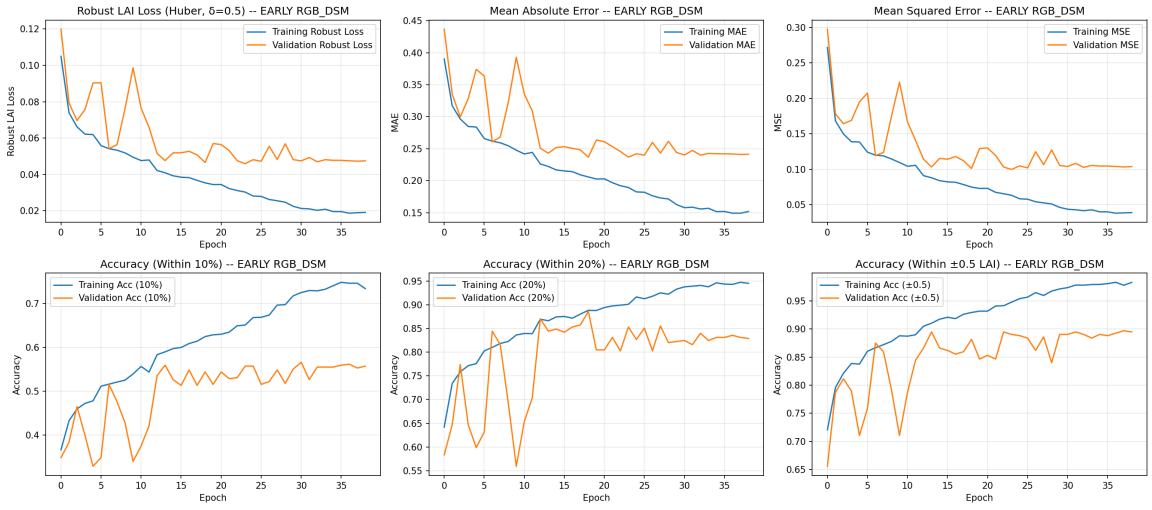


Fig. 4.3: Training and Validation Curves for Early RGB-DSM combination

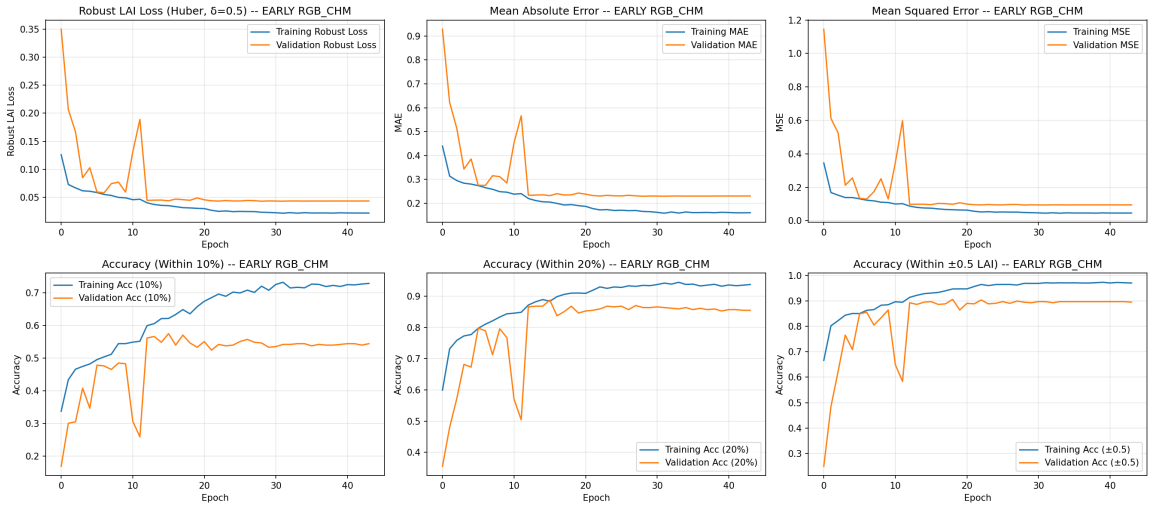


Fig. 4.4: Training and Validation Curves for Early RGB-CHM combination

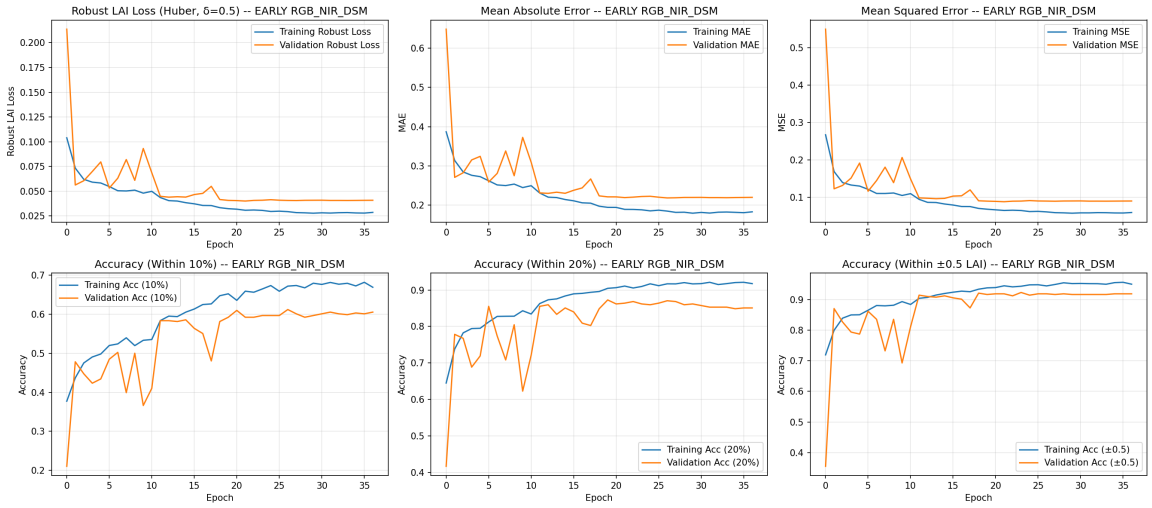


Fig. 4.5: Training and Validation Curves for Early RGB-NIR-DSM combination

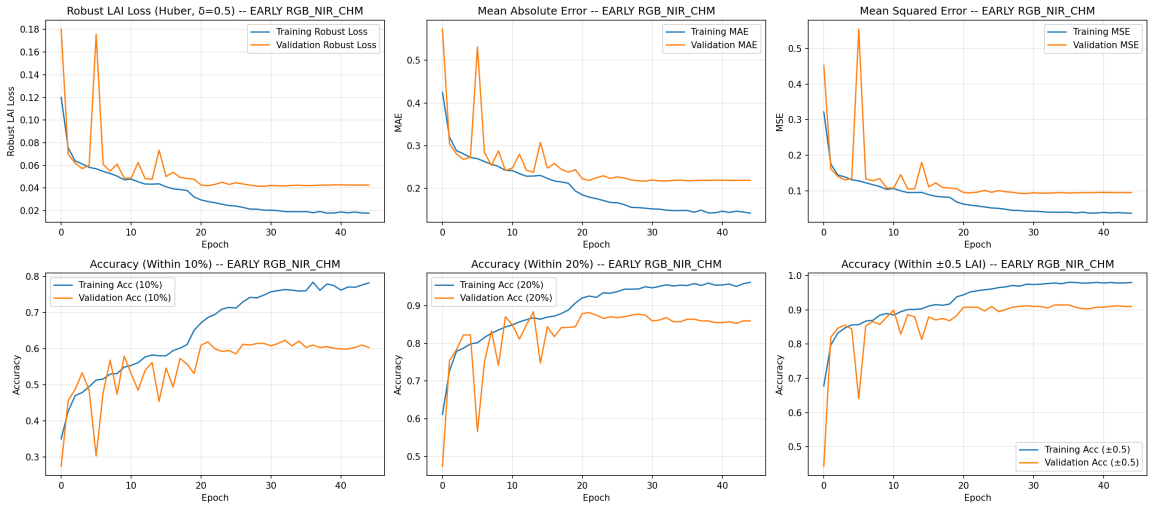


Fig. 4.6: Training and Validation Curves for Early RGB-NIR-CHM combination

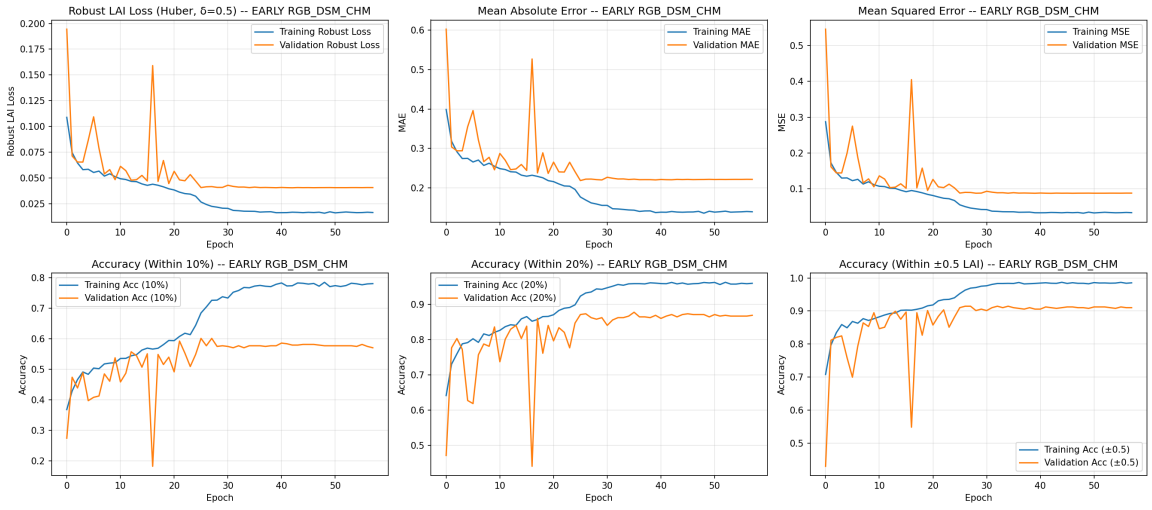


Fig. 4.7: Training and Validation Curves for Early RGB-DSM-CHM combination

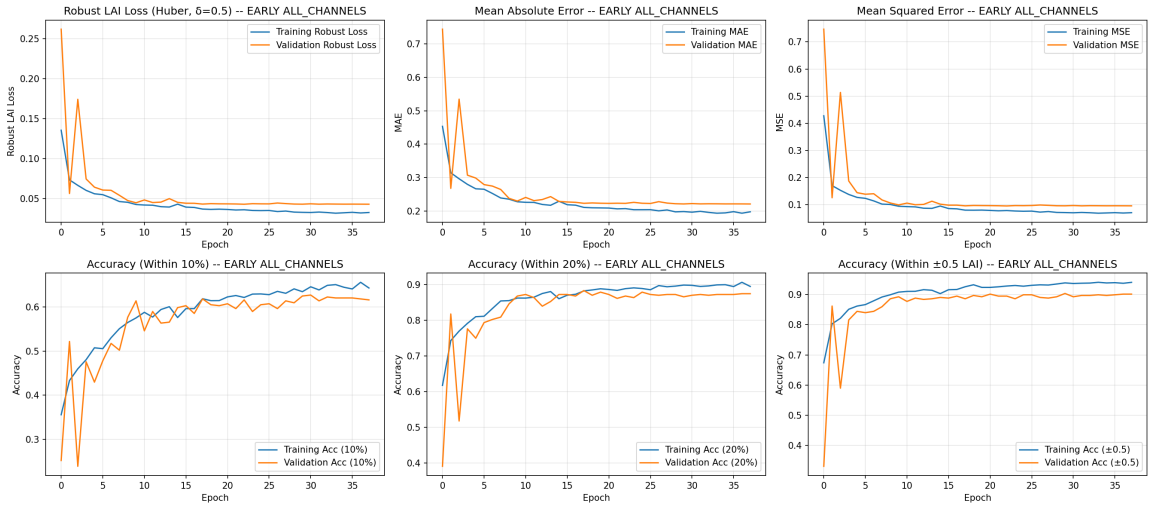


Fig. 4.8: Training and Validation Curves for Early All Channels combination

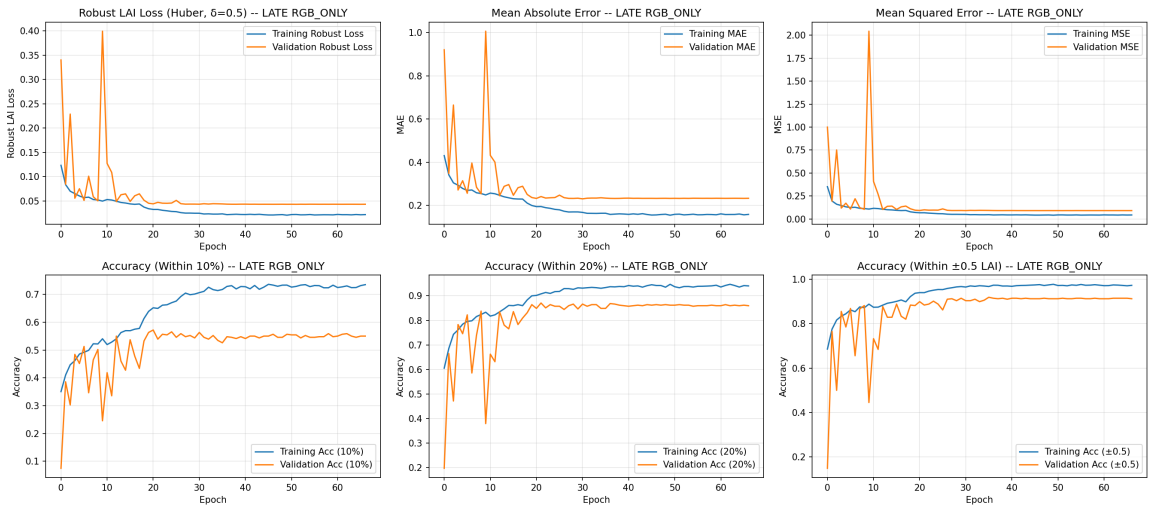


Fig. 4.9: Training and Validation Curves for Late RGB combination

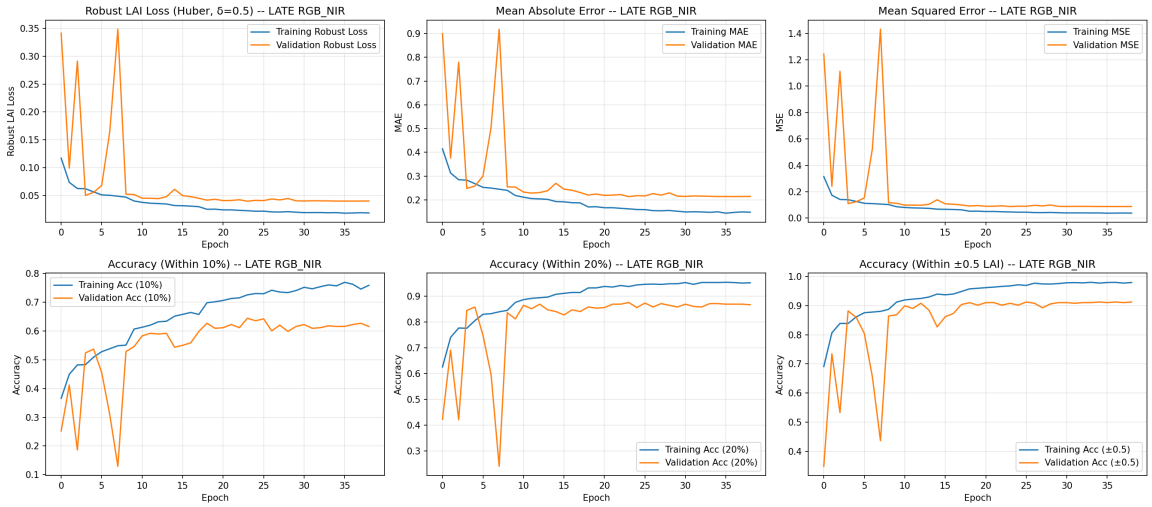


Fig. 4.10: Training and Validation Curves for Late RGB-NIR combination

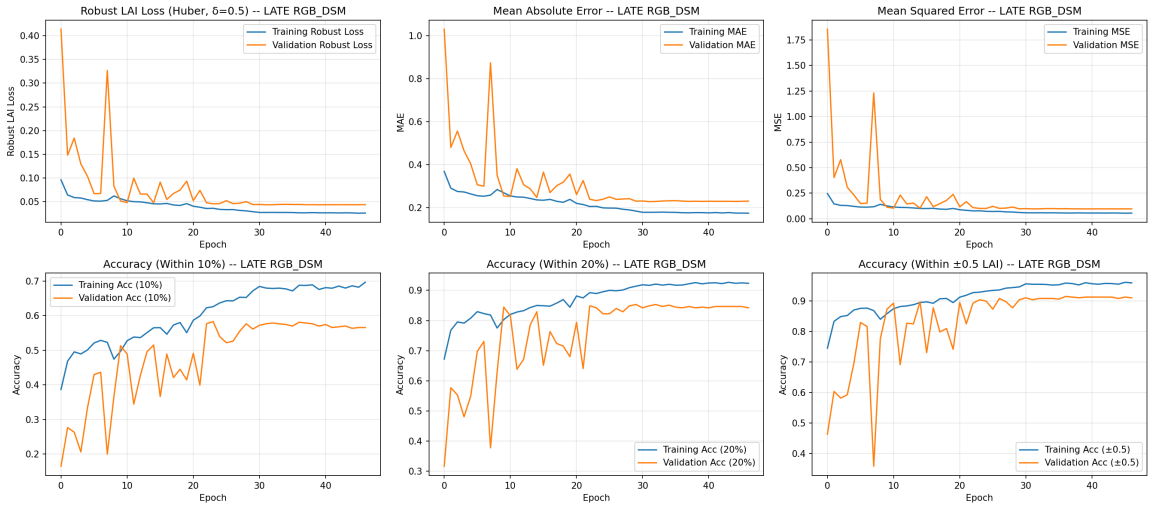


Fig. 4.11: Training and Validation Curves for Late RGB-DSM combination

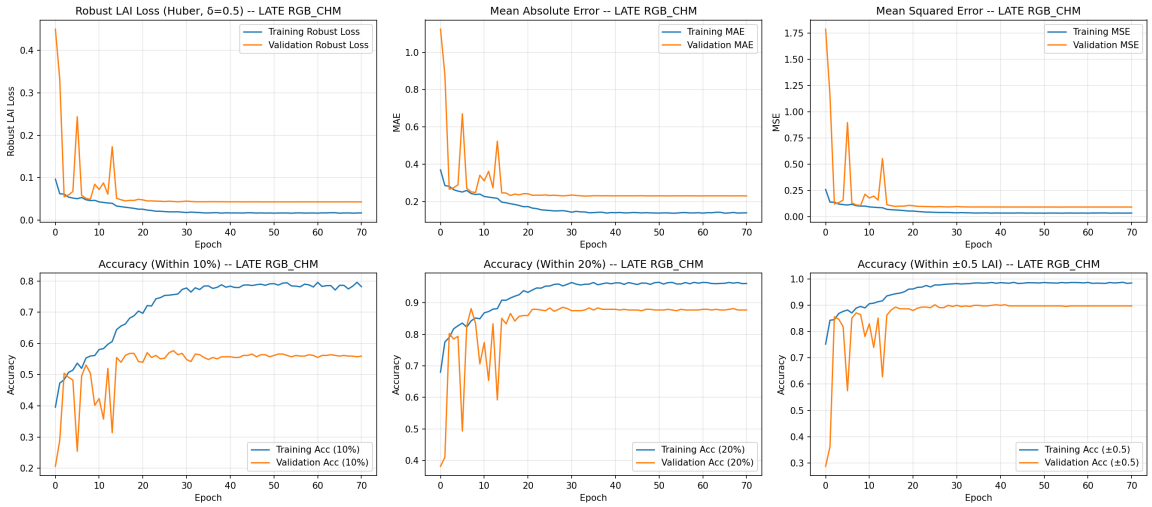


Fig. 4.12: Training and Validation Curves for Late RGB-CHM combination

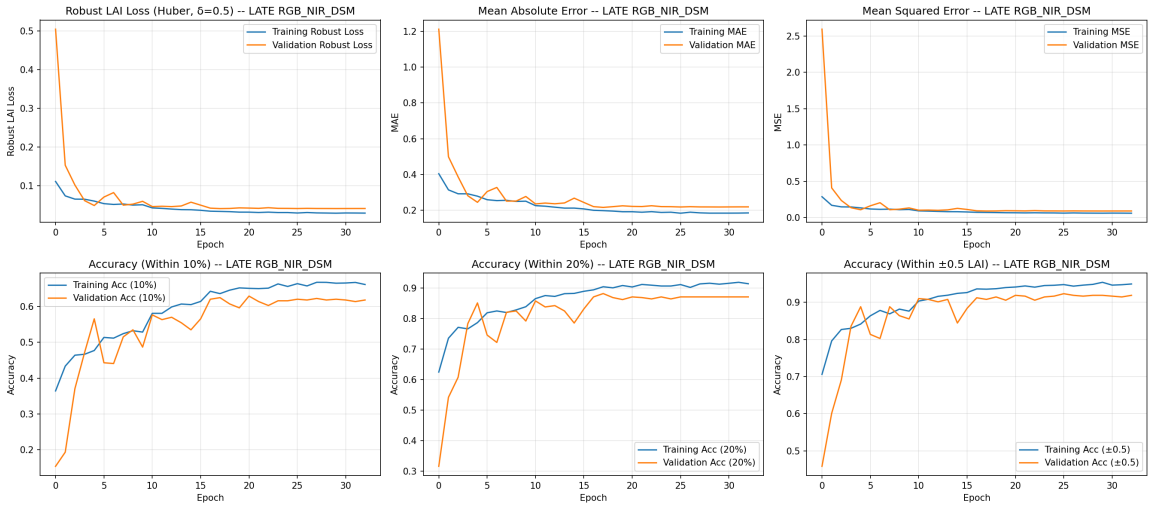


Fig. 4.13: Training and Validation Curves for Late RGB-NIR-DSM combination

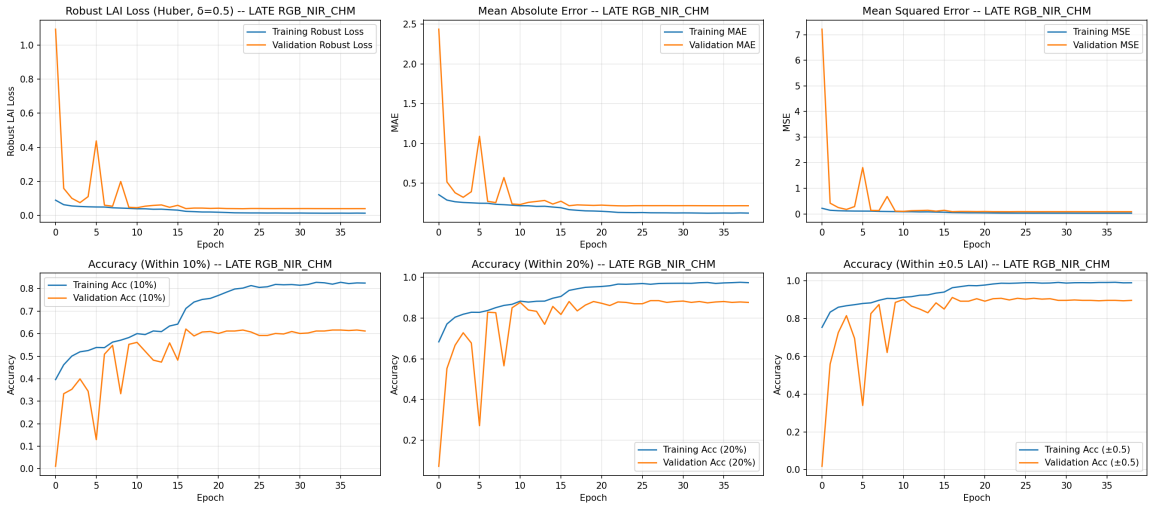


Fig. 4.14: Training and Validation Curves for Late RGB-NIR-CHM combination

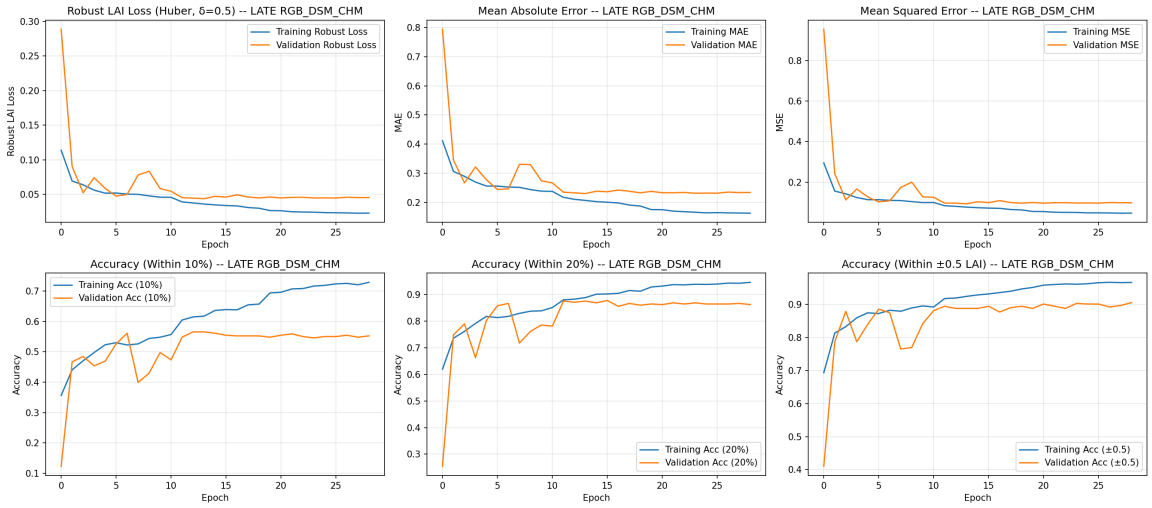


Fig. 4.15: Training and Validation Curves for Late RGB-DSM-CHM combination

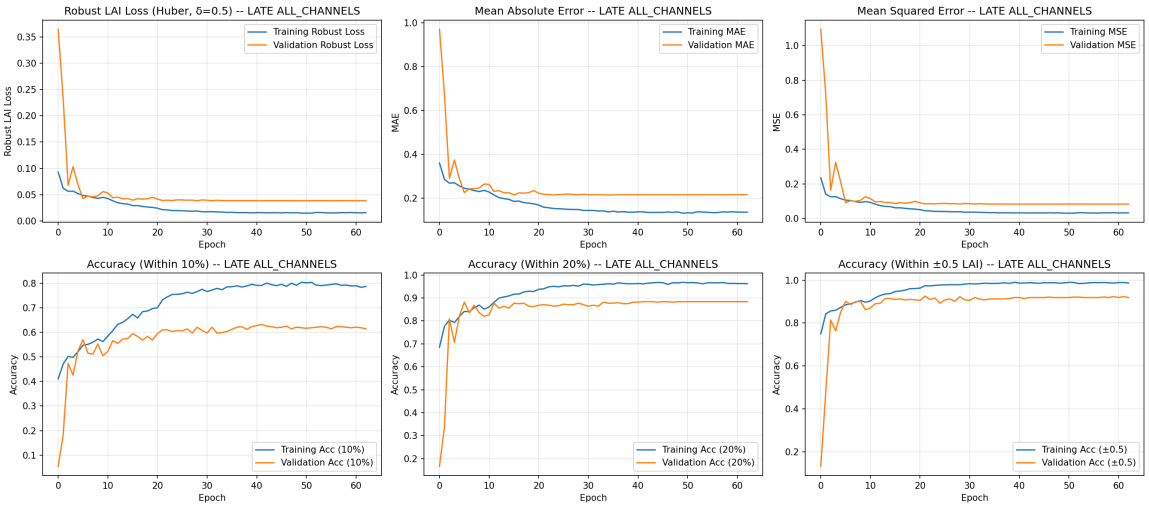


Fig. 4.16: Training and Validation Curves for Late All Channels combination

get closer together than the accuracy lines do. Accuracy within 10 percent, the toughest threshold, almost always shows the largest and most stubborn gap between training and validation. In RGB-NIR-CHM under Early Fusion (Figure 4.6), for instance, training and validation MAE track closely throughout the run, yet training accuracy within 10% climbs to nearly 78% while validation plateaus around 60%, a gap the loss curves alone would not reveal.

RGB-DSM under Early Fusion (Figure 4.3) stands out from every other Early Fusion model as its training loss, MAE, and MSE never plateau, continuing to decrease for the full length of the run, while the corresponding validation curves stop improving after roughly epoch 15 and its training accuracy keeps climbing well past where validation has flattened, producing the widest training-validation gap of any Early Fusion model. This is consistent with RGB-DSM Early being one of the two weakest models on the held-out test set (Section 4.1). Under Late Fusion, the same channel combination (Figure 4.11) shows a comparatively narrower gap between training and validation accuracy, mirroring the test-set result that Late Fusion outperforms Early Fusion specifically for RGB-DSM.

The three models that trained the longest before early stopping was triggered, RGB Only, All Channels, and RGB-CHM under Late Fusion (Figures 4.9, 4.16, and 4.12, each run for roughly 60–70 epochs, compared with under 45 epochs for their Early Fusion coun-

terparts in Figures 4.1, 4.8, and 4.4), also end training with the widest training-validation accuracy gaps observed among all 16 models. The training accuracy within 10% climbs to around 73–80%, but validation stays lower, between 55–62%. This matches what's seen in Section 4.1. Although Late Fusion uses up to nearly four times as many parameters as Early Fusion (Table 3.2), that extra complexity doesn't improve performance on unseen test data. When models train for a long time and the gap between training and validation accuracy gets bigger, it usually means the model is getting better at memorizing the training set but not actually getting better at handling new data. This tendency shows up with Late Fusion in these three combinations, but it's not something you see in every Late Fusion model.

This overfitting signature, however, is not universal across every Late Fusion model. RGB-NIR-DSM and RGB-DSM-CHM under Late Fusion (Figures 4.13 and 4.15) trained for only 33 and 29 epochs respectively, the shortest of all 16 runs, and both show comparatively tight training-validation tracking in their accuracy curves rather than a widening gap. Yet neither combination ranks among the stronger models on the held-out test flights (Section 4.1); RGB-NIR-DSM Late in particular is one of the weaker performers overall. Since the validation split is drawn from the same pooled flights as training (Section 3.5), while the test flights are entirely separate UAS flights on different dates, tight training-validation tracking here reflects good generalization within the same flights the model was trained on, but not necessarily to a genuinely new flight. Training and validation curves show how well a model is learning from familiar data, but they don't tell the whole story. To really know if a model will hold up in practice, it's important to see how it performs on completely new data, just like what's done in the held-out test evaluation in Section 4.1.

Beyond explaining Late Fusion's weaker held-out accuracy, this overfitting signature has a direct practical implication for how to deploy these models. A model that has overfit to the pooled training and validation flights can still appear reliable if judged only by validation performance, since validation chips are drawn from the same flights the model was trained on and therefore share their site-specific conditions, illumination, and canopy structure; the risk only becomes visible once the model is applied to a genuinely new flight,

as the held-out test evaluation in Section 4.1 and the site-level breakdown in Section 4.3 do, and as the field-scale maps in Chapter 5 do at an even larger scale. This is precisely why every model in this thesis is evaluated on two held-out test flights never seen during training or validation, rather than on validation performance alone, and why generalization to an entirely new flight, not validation accuracy, is treated throughout Section 4.3 and Chapter 5 as the standard a model must meet to be considered reliable for field-scale or operational use.

4.3 Test Performance Analysis on BAR and RIP Datasets

Sections 4.1 and 4.2 pooled both held-out test flights, BAR_190627_1207 and RIP_180619_1120, together. Splitting the same predictions out by site (Figure 4.17) shows that the two flights behave very differently as BAR chips ($n = 340$ per combination, actual LAI roughly 0.8–3.2) reach a moderate fit (Early $R^2 = 0.45$, Late $R^2 = 0.36$), while RIP chips ($n = 170$ per combination, actual LAI roughly 2.0–4.8) barely correlate with actual LAI at all (Early $R^2 = 0.09$, Late $R^2 = 0.08$). This site-level gap holds regardless of which channels are used, so it is examined for each channel combination individually below.

On BAR (Figures 4.18–4.25, left panels), several combinations reach R^2 around 0.6 under Early Fusion: RGB-CHM ($R^2 = 0.62$, MAE= 0.20, RMSE= 0.26), All Channels ($R^2 = 0.61$, MAE= 0.21, RMSE= 0.27), RGB-NIR-DSM ($R^2 = 0.61$, MAE= 0.21, RMSE= 0.27), and RGB-NIR-CHM ($R^2 = 0.61$, MAE= 0.23, RMSE= 0.29). Early Fusion matches or beats its Late Fusion counterpart on BAR for six of the eight combinations; the two exceptions are RGB-NIR and All Channels, where Late Fusion is marginally ahead ($R^2 =$ 0.50 vs. 0.46, and $R^2 = 0.62$ vs. 0.61, respectively). RGB alone (Early $R^2 = 0.31$, Late $R^2 = 0.26$) and RGB-DSM (Early $R^2 = 0.25$, Late $R^2 = 0.15$) are clearly the weakest combinations on BAR under either fusion strategy, consistent with DSM alone adding little useful signal (Section 4.1).

On RIP (Figures 4.18–4.25, right panels), every combination and fusion strategy collapses to a weak fit, with R^2 ranging from just 0.04 (RGB-NIR-DSM, Early) to 0.16 (RGB-CHM, Early, the best of all sixteen on this site). In every panel, the trend line is almost

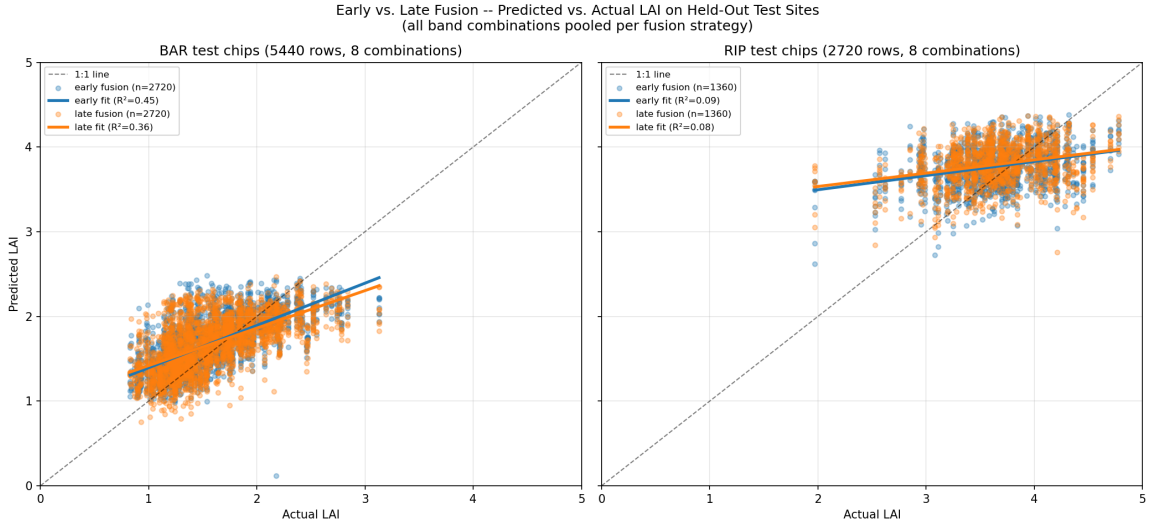


Fig. 4.17: Predicted vs. actual LAI on the two held-out test flights, BAR (left) and RIP (right), pooled across all eight channel combinations, for Early and Late Fusion

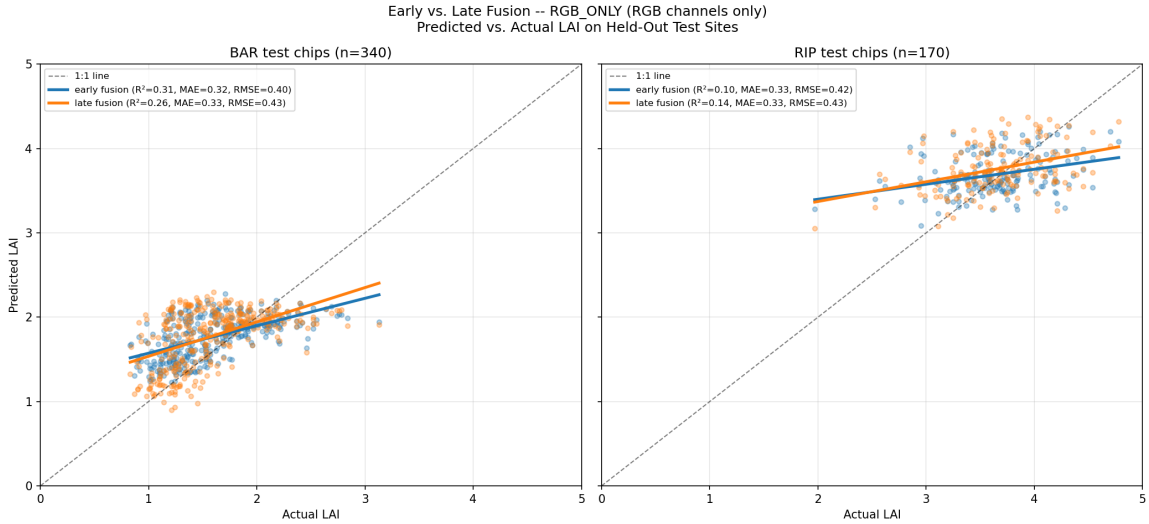


Fig. 4.18: Scatter Plot for Test Results on BAR and RIP images by RGB-only combination model

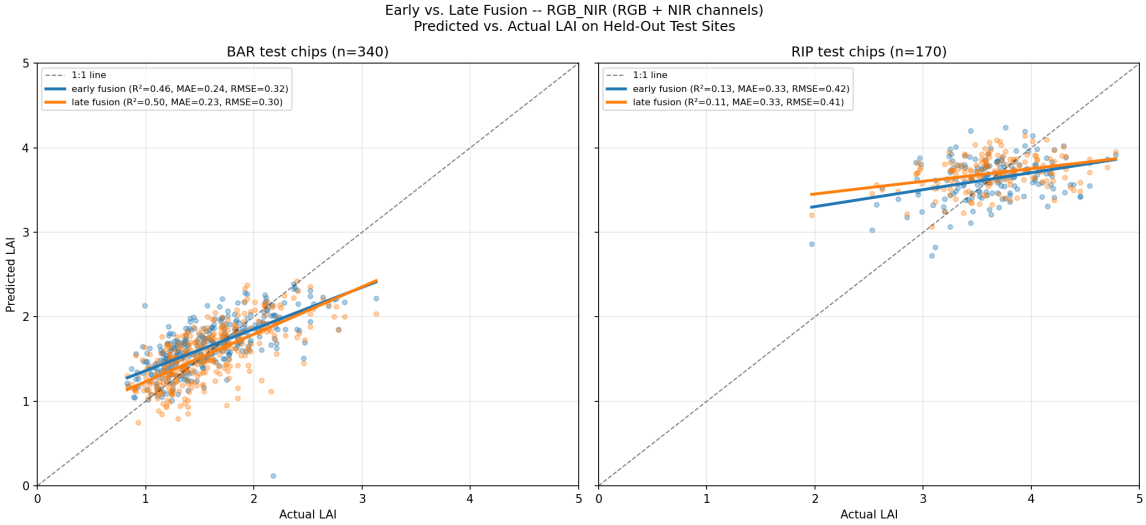


Fig. 4.19: Scatter Plot for Test Results on BAR and RIP images by RGB-NIR combination model

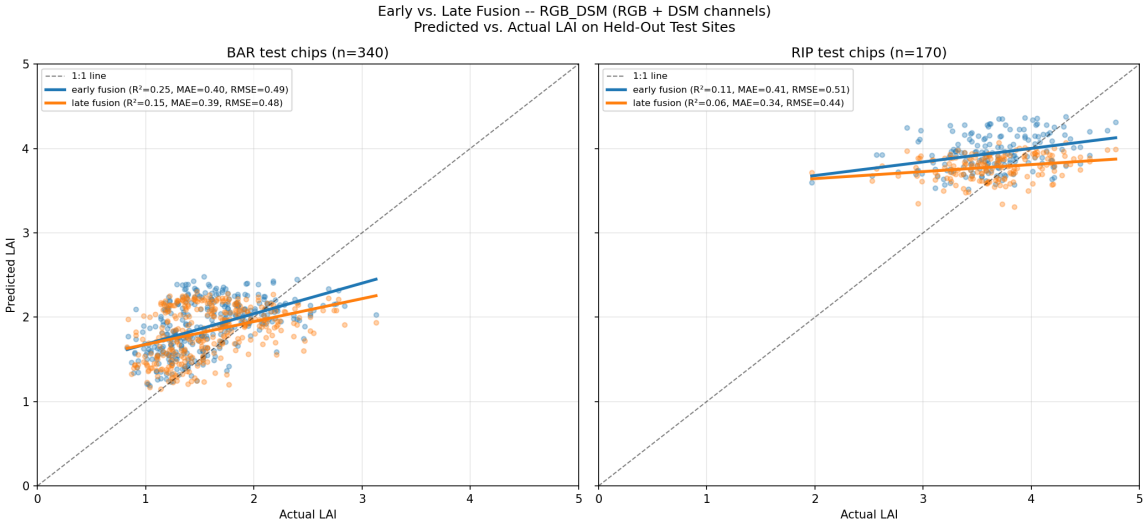


Fig. 4.20: Scatter Plot for Test Results on BAR and RIP images by RGB-DSM combination model

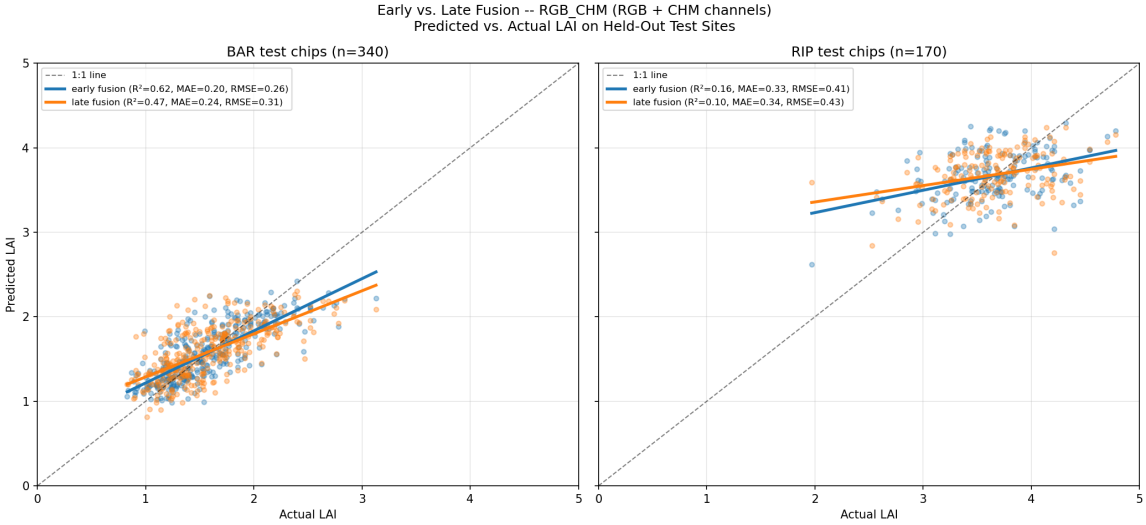


Fig. 4.21: Scatter Plot for Test Results on BAR and RIP images by RGB-CHM combination model

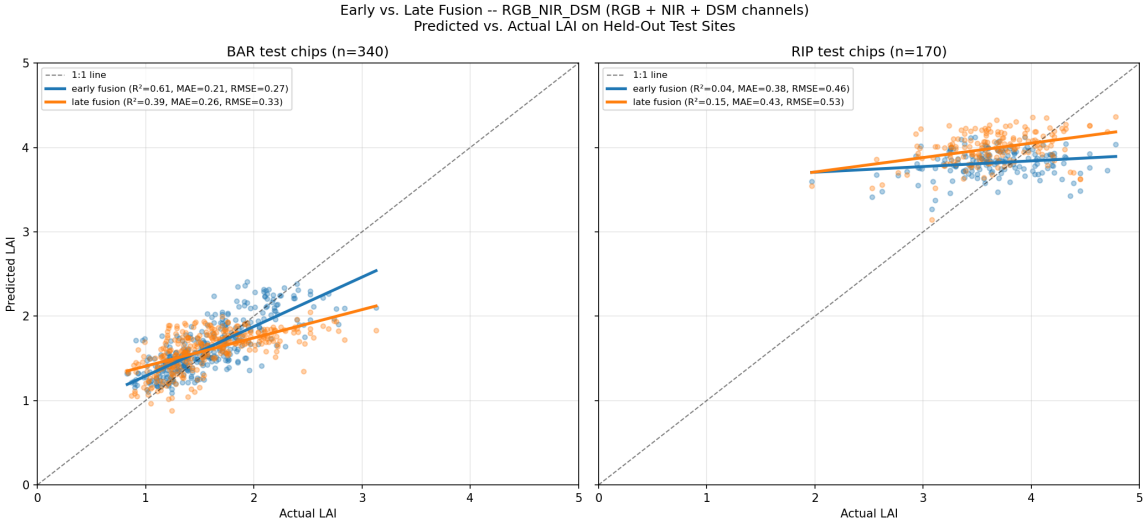


Fig. 4.22: Scatter Plot for Test Results on BAR and RIP images by RGB-NIR-DSM combination model

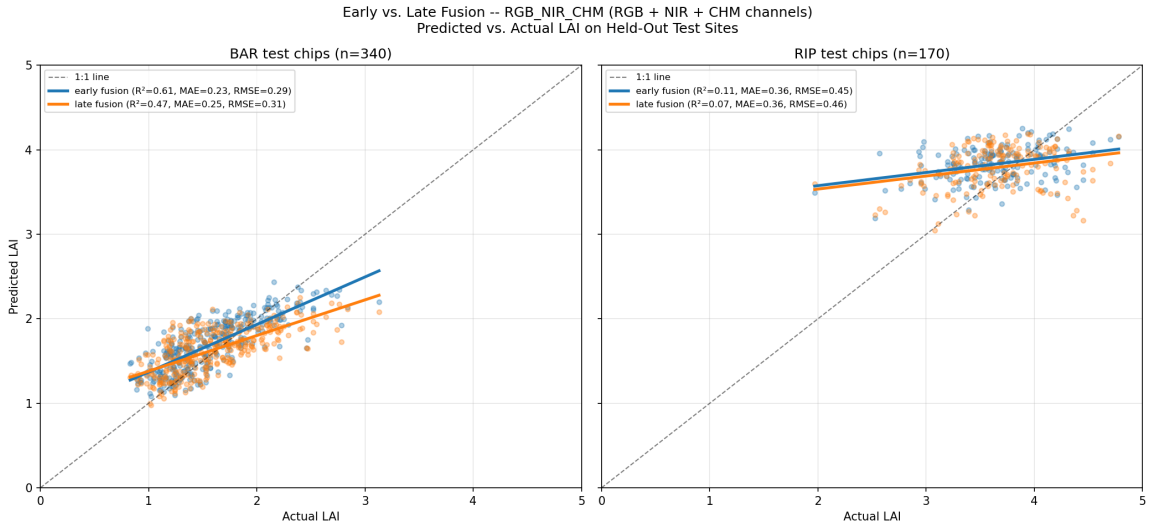


Fig. 4.23: Scatter Plot for Test Results on BAR and RIP images by RGB-NIR-CHM combination model

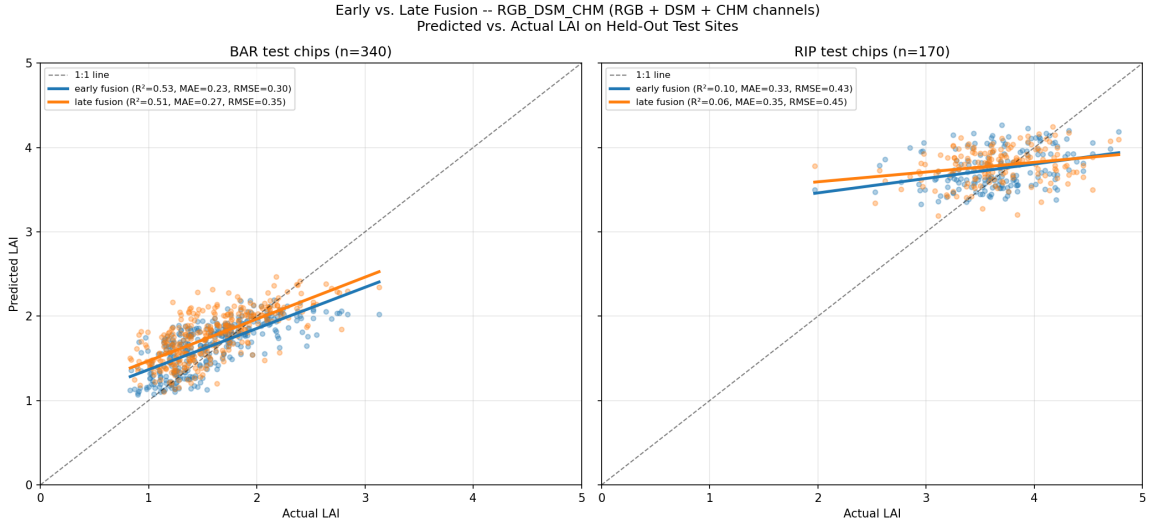


Fig. 4.24: Scatter Plot for Test Results on BAR and RIP images by RGB-DSM-CHM combination model

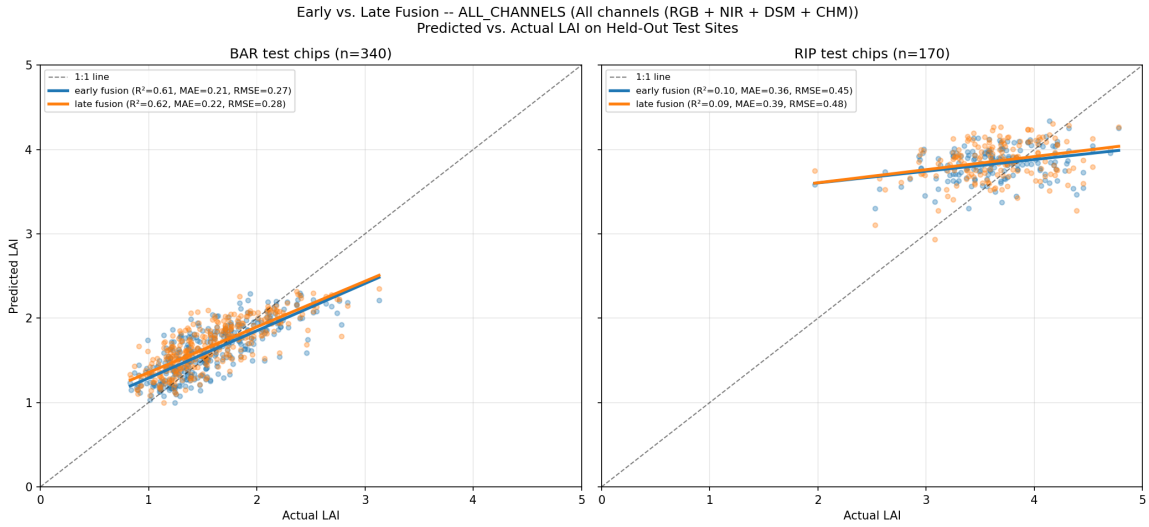


Fig. 4.25: Scatter Plot for Test Results on BAR and RIP images by All Channels combination model

flat. Most of the predicted LAI values fall within a narrow band, roughly 3.5 to 4.0, even when the actual LAI ranges from around 2.0 up to 4.8. This means the models aren't really telling apart the lower and higher LAI chips in this flight. Notably, MAE and RMSE on RIP (typically 0.33–0.51) are not dramatically worse in absolute LAI units than on BAR (typically 0.20–0.40); the near-zero R^2 values are driven as much by RIP's narrower, higher actual-LAI range compressing the denominator of R^2 as by a genuine loss of predictive power, and are also consistent with the known tendency of the underlying indirect PCA measurement itself to saturate and underestimate at high LAI (Section 3.2).

Taken together, the by-site comparison shows that channel combination and fusion strategy matter far more for a moderate-density flight like BAR, where the choice between the weakest combination (RGB-DSM, $R^2 \approx 0.15$–0.25) and the strongest (RGB-CHM or All Channels, $R^2 \approx 0.6$) changes the fit substantially, than for a dense-canopy flight like RIP, where no combination or fusion strategy meaningfully explains the variation in actual LAI. This suggests that improving generalization to RIP-like, high-LAI conditions is unlikely to be solved by choosing a different channel combination alone, and instead points to a more fundamental limitation, whether in the range of LAI values represented in training, or in the saturation behavior of the underlying ground-truth measurement itself at high canopy

density.

CHAPTER 5

INFERENCE ON A COMPLETE UAS FLIGHT DATA AND SATELLITE DATA

5.1 Field-Scale LAI Mapping on a whole UAS flight

Although the solution in the previous section received some realistic and satisfactory results based on the goal it was trying to achieve (Table 4.2, 4.1), i.e., proposing a pipeline to prove which combinations of data can give the perfect prediction based on the LAI value range, it needs to be further analyzed on a specific site to understand which combination of channels will have better prediction estimates.

5.1.1 Field-Scale Inference Pipeline

To evaluate the trained models beyond the individual, isolated chips used for training and testing, each of the 16 models (Section 3.6) was applied across an entire UAS flight to produce a full field-scale LAI map, following the pipeline summarized in Figure 5.1.

Rather than feeding a whole flight's orthomosaic into a model directly, the orthomosaic was first cut into the same 25×25 pixel chip grid used during training, one full grid per channel combination, so that every model could be applied to input chips of the size it was trained on. At the dataset's 15 cm pixel^{-1} resolution (Chapter 3), each 25×25 px chip covers a 3.6 m $\times$ 3.6 m (12.96 m^2) ground footprint, with each individual pixel representing a 15 cm $\times$ 15 cm (0.0225 m^2) patch of the field. Each combination's grid held over 15,000 chips covering the full extent of the flight, particularly chosen from one of the RIP flights (Figure 5.2). Every chip in a combination's grid was pre-processed identically to training, using the same global z-score normalization and the same cached training-set channel statistics described in Section 3.4, so that predictions at inference time would be consistent with what each model was trained on.

Because each trained model outputs a single LAI value per 25×25 chip rather than a

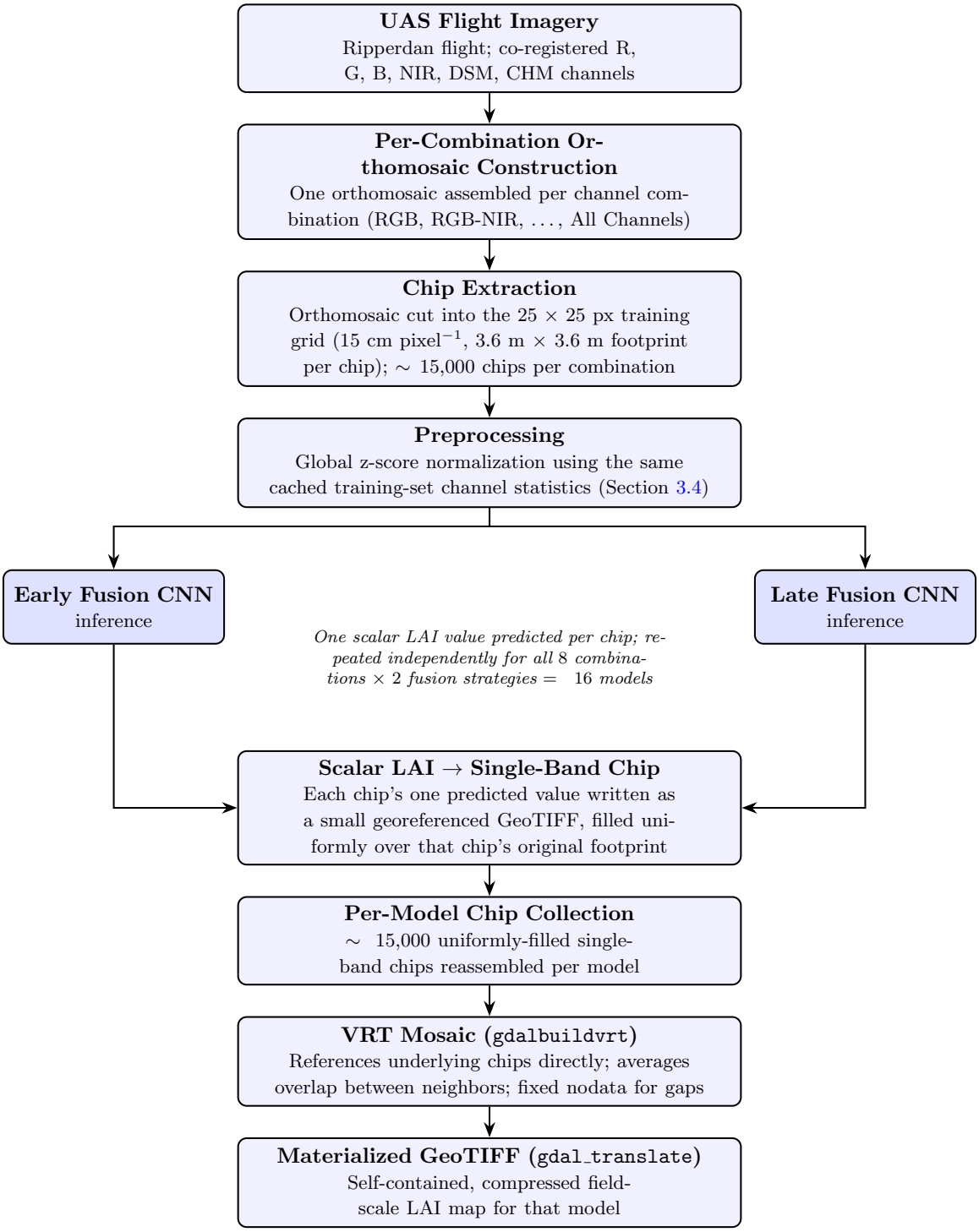


Fig. 5.1: Field-scale inference pipeline: an orthomosaic is built per channel combination from the target UAS flight, cut into the same 25 × 25 px (15 cm pixel^{-1}, 3.6 m × 3.6 m footprint) chip grid used in training, and passed through each trained model to produce one scalar LAI value per chip. Each value is written back out as a uniformly-filled single-band chip, and the resulting chip collection is assembled into a VRT mosaic and materialized into a single field-scale GeoTIFF, independently for all 16 models.

per-pixel prediction, applying a model to a chip grid produces one scalar per chip. To turn these scalar predictions back into a spatial map, each chip's predicted LAI value was written out as a small georeferenced single-band GeoTIFF, filled uniformly with that value, using the coordinate reference system and spatial transform copied from the chip's corresponding location in the source grid so that every prediction remained correctly geolocated. The resulting collection of single-band chips for a given model, again on the order of 15,000 files, was then assembled into one continuous mosaic. An initial attempt was made to load the full set of individual chip GeoTIFFs generated by the model directly into GIS software (QGIS). However, the sheer number of files repeatedly caused the software to crash, which created issues in generating the complete LAI distribution map, making the predicted maps unusable for visualization or further analysis. To resolve this, a Virtual Raster (VRT) was built instead. It is a lightweight GDAL format that references the underlying chip files rather than duplicating their pixel data, using an averaging rule to resolve any overlap between neighboring chips, and a fixed nodata value to mark locations with no chip present. The VRT was then materialized into a single self-contained, compressed GeoTIFF, so that the final field-scale LAI map did not depend on the underlying per-chip files remaining in place and could be opened directly as one raster layer without crashing the GIS software. This chip generation, prediction, and mosaicking process was repeated independently for all 16 models, producing the field-scale LAI maps shown in Figures 5.3 and 5.4.

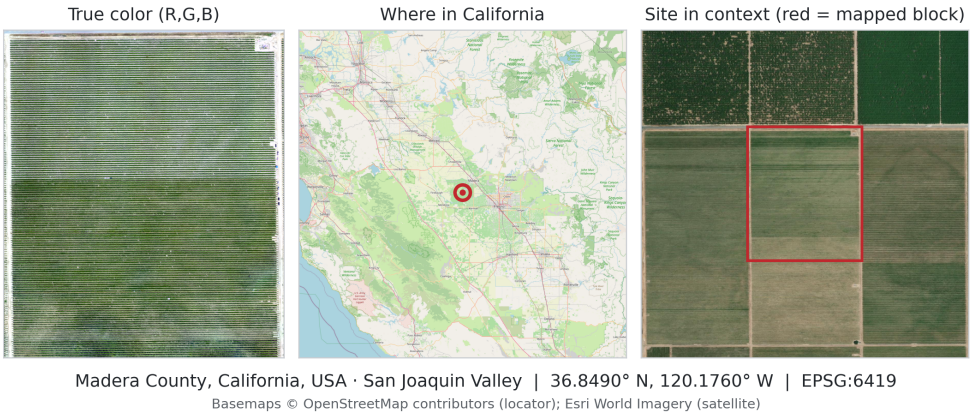


Fig. 5.2: Study site: UAS true-colour orthomosaic of the mapped block (4 May 2019, 0.15 m GSD; contrast-stretched for display), regional locator, and satellite context with the block outlined in red

5.1.2 Predicted LAI Maps Across Channel Combinations

Across the eight panels of Figure 5.3 (Early Fusion), the vine-row structure of the block is visible in every combination as fine horizontal banding, but the overall level and contrast of predicted LAI differs sharply depending on whether DSM is one of the active channels. RGB Only, RGB-NIR, RGB-CHM, and RGB-NIR-CHM all produce pale, low-contrast maps in which most of the block stays within the low-to-moderate LAI range and row-to-row differences are muted. RGB-DSM, RGB-NIR-DSM, and RGB-DSM-CHM, by contrast, all produce visibly darker maps with pronounced dark-green horizontal banding reaching much higher predicted LAI values across large parts of the block. All Channels falls between these two groups, showing moderate, fairly uniform row banding without the extreme saturation of the DSM-containing combinations or the washed-out appearance of the others. RGB-CHM and RGB-NIR-CHM additionally show a scatter of isolated, near-black single-chip outliers with no corresponding row structure, consistent with occasional noisy or out-of-distribution chips rather than genuine spatial signal. Across all eight Early Fusion combinations, the models tend to pick up the unpaved road on the block's left edge as an area with almost no LAI. As a result, the road blends into the light background on the map and doesn't stand out as a separate feature, mirroring how it looks in real life.

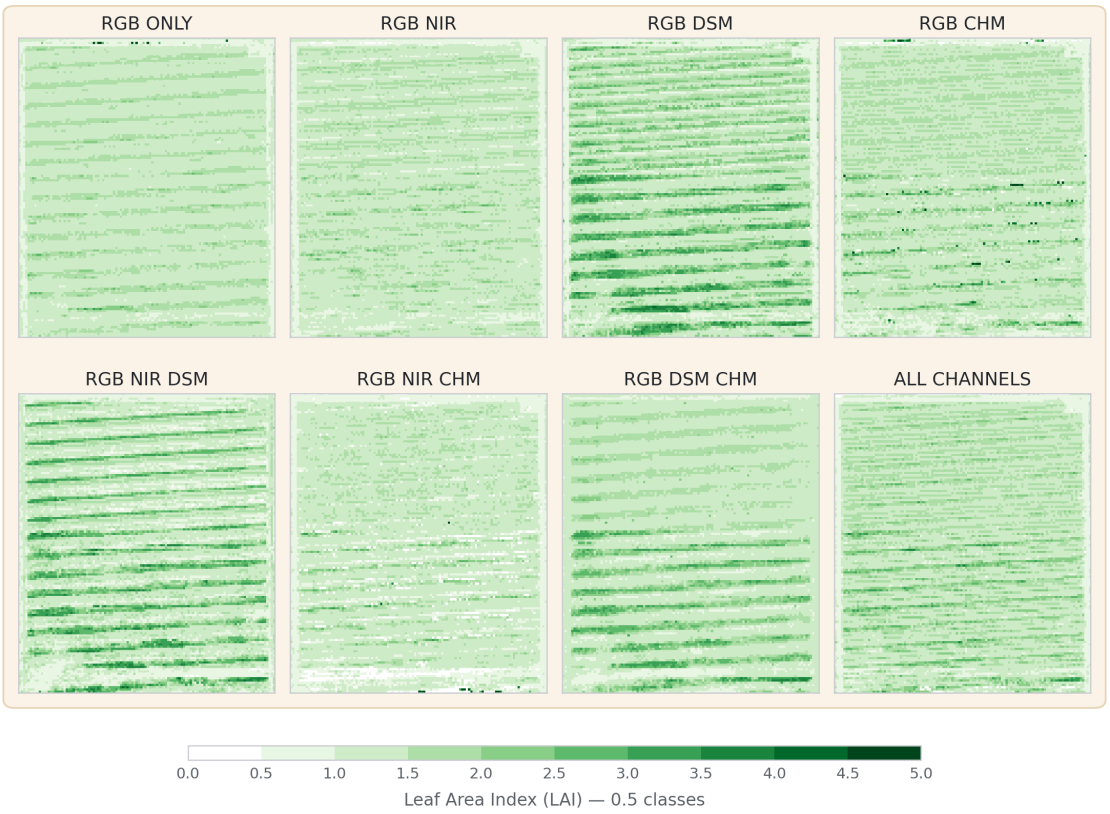


Fig. 5.3: Predicted LAI under **Early Fusion** for eight channel combinations

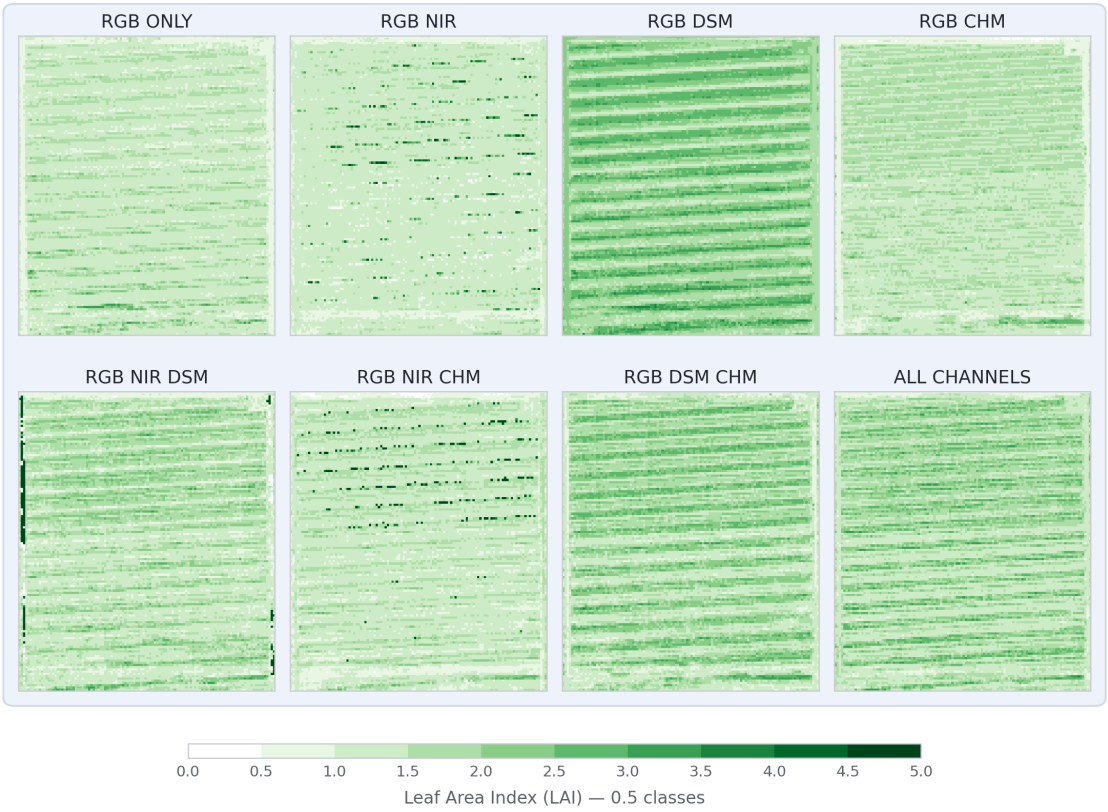


Fig. 5.4: Predicted LAI under **Late Fusion** for eight channel combinations

The true-colour context image (Figure 5.2) also shows the mapped block splitting into two visually distinct halves across its horizontal center, a sparser canopy in the top half and a markedly denser canopy in the bottom half; of the eight Early Fusion combinations, only RGB-DSM and RGB-DSM-CHM clearly reproduce this top-to-bottom density split in their predicted LAI, while the rest show comparatively uniform row banding from top to bottom.

Figure 5.4 (Late Fusion) shows the same DSM-driven split, and in some respects more strongly. RGB-DSM's dark banding now extends across effectively the entire mapped block rather than being concentrated in part of it; this block-wide over-prediction also swallows the unpaved access road along the block's left edge, which no longer stands out as a distinct feature the way it does under Early Fusion. RGB-DSM-CHM shows a similarly strong, block-wide pattern. RGB-NIR and RGB-NIR-CHM both show a dense scatter of isolated dark-chip artifacts spread across most of the block, more widespread than the sparser outliers seen under Early Fusion for the CHM-containing combinations. RGB-NIR-DSM shows a solid, dark-green vertical stripe running the full height of the block along its left edge; this corresponds to the same unpaved road, correctly picked out here as a distinct linear feature but mispredicted as high rather than near-zero LAI. RGB-DSM and RGB-NIR-DSM are therefore the only two of the eight Late Fusion combinations that fail to reproduce the road's low LAI seen consistently across every Early Fusion map, a genuine failure mode rather than a feature of the canopy itself. As under Early Fusion, All Channels produces the most visually consistent and artifact-free map of the eight combinations.

The combinations that produce the most visually extreme, high-contrast full-flight maps here, RGB-DSM above all, are the same combinations identified in Sections 4.1 and 4.3 as the weakest and least consistent on the held-out test chips. This agreement suggests that DSM's poor chip-level accuracy is not limited to a small number of difficult test chips. Instead, this pattern shows up across the entire flight. Models that include DSM tend to predict higher LAI values and make the differences between vine rows more noticeable all across the block, not just in a few places. All Channels, despite not being the single best performer on the held-out test chips in either fusion strategy (Section 4.1), produces

the most spatially coherent full-flight map under both fusion strategies, with neither the saturation seen in the DSM-containing combinations nor the speckled artifacts seen in RGB-NIR and RGB-NIR-CHM.

Beyond their level and contrast, the DSM-containing combinations' banding also differs from the true vine-row structure in its orientation. In the true-colour reference image (Figure 5.2), the vine rows run essentially flat and horizontal across the block, and the banding reproduced by the RGB-, CHM-, and NIR-only combinations tracks this orientation closely. In RGB-DSM, RGB-NIR-DSM, and RGB-DSM-CHM, however, the dark banding is visibly tilted, drifting diagonally across the panel rather than running flat, an angle not present in the true row structure. This is consistent with a well-documented failure mode of UAS structure-from-motion photogrammetry: without an extensive network of ground control points, UAS-derived digital surface models are known to exhibit a smooth, systematic tilt or "doming" error arising from the near-parallel imaging geometry of a single flight combined with imperfectly-corrected radial lens distortion in the self-calibrating bundle adjustment [44]. Because this systematic error is governed by the flight's own imaging geometry rather than by the vine-row planting direction, it would be expected to appear as banding at an angle to the true rows, precisely the pattern visible here, and it plausibly also explains why the artifact is concentrated in the DSM-containing combinations and largely absent from CHM-only and RGB-only maps: CHM is computed as DSM minus a separately-derived DTM (Section 3.3), which would at least partially cancel a systematic elevation error common to both surfaces, whereas DSM carries it directly into the model's input.

5.1.3 Quantitative Summary of Predicted LAI Across the Flight

The qualitative patterns described above are reflected directly in the pooled per-model statistics computed over the full set of 15,162 chips tiling the mapped Ripperdan flight (Figure 5.5). Because each trained model outputs a single scalar LAI value per chip rather than a per-pixel prediction (Section 5.1), pooling every chip's prediction for a given model yields a mean, standard deviation, and range that summarize that model's behavior across

the entire flight, independent of any spatial pattern.

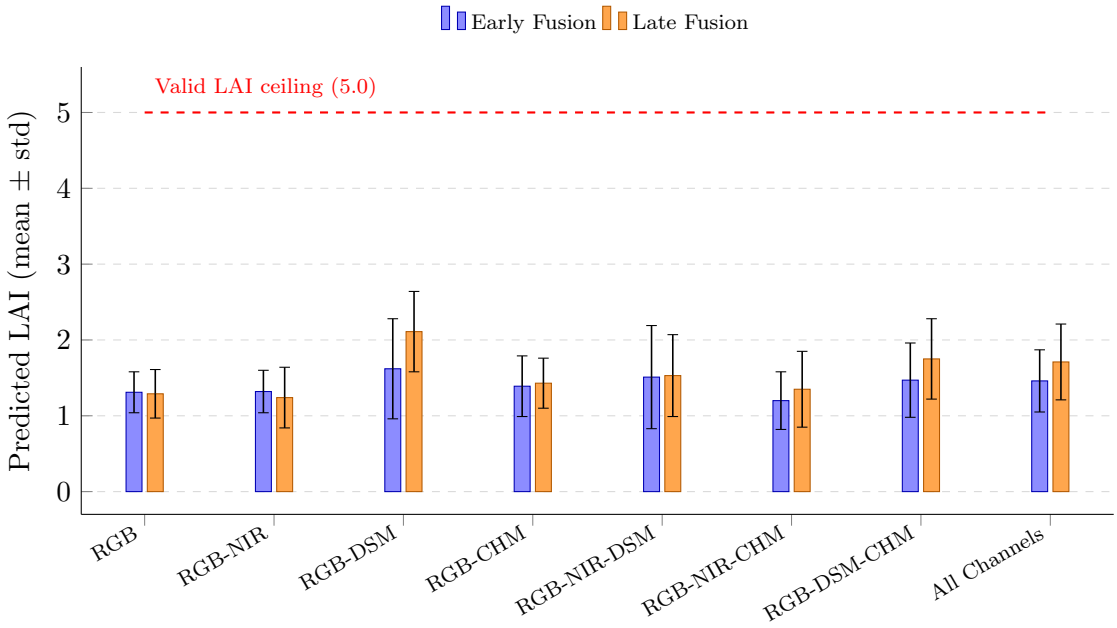


Fig. 5.5: Predicted LAI (mean ± std across all 15,162 chips) for each of the 16 trained models applied to the full Ripperdan flight, by channel combination and fusion strategy.

Averaged across all 15,162 chips, every one of the 16 models predicts a mean LAI between 1.20 and 2.11, comfortably inside the physically valid $[0, 5]$ range enforced by the sigmoid output layer (Section 3.6) and consistent with a site that is predominantly low-to-moderate canopy rather than uniformly dense. Within that range, however, the DSM-containing combinations again stand apart. RGB-DSM has the highest mean of any model under Late Fusion (2.11), with RGB-DSM-CHM and All Channels close behind as the second- and third-highest Late Fusion means (1.75 and 1.71), while RGB-NIR-DSM has the largest standard deviation of any model under either fusion strategy (0.68 Early, 0.54 Late). Every non-DSM combination, by contrast, has a mean between 1.20 and 1.43 under both fusion strategies, a substantially narrower band. This mirrors the qualitative description above that the DSM-containing models do not merely produce visibly darker maps, they systematically shift the entire flight's predicted LAI distribution upward and widen its spread.

The gap between fusion strategies is itself concentrated in specific channel combinations rather than spread uniformly across every DSM-containing one. For every combination without DSM, Early and Late Fusion means differ by no more than 0.15 LAI units (e.g., RGB-CHM: 1.39 Early versus 1.43 Late; RGB-NIR-CHM: 1.20 versus 1.35, the largest non-DSM gap). RGB-DSM, RGB-DSM-CHM, and All Channels show substantially larger gaps, with Late Fusion's mean exceeding Early Fusion's by 0.49 (2.11 vs. 1.62), 0.28 (1.75 vs. 1.47), and 0.25 (1.71 vs. 1.46) respectively, a quantitative confirmation of the pattern visible in Figure 5.4, where RGB-DSM's dark banding under Late Fusion extends across the entire block rather than the partial coverage seen under Early Fusion. RGB-NIR-DSM is the exception among the DSM-containing combinations, its Early and Late means differ by only 0.02 (1.51 vs. 1.53), despite having the largest standard deviation of any model overall. Rather than producing a uniform increase across all values, this combination tends to create high variance due to a handful of extreme chips. These outlier values are often linked to mispredictions along the road, as seen in Late RGB-NIR-DSM, while a few other chips display the opposite effect by reaching near-zero LAI values.

Several combinations, most often but not exclusively those with CHM as an active channel, include at least one chip whose predicted LAI saturates at or very near the sigmoid's 0 or 5 bound. Early RGB-CHM ($[0.00, 4.94]$), Early RGB-NIR-CHM ($[0.00, 5.00]$), Late RGB-NIR-CHM ($[0.10, 5.00]$), and Late RGB-NIR-DSM ($[0.00, 5.00]$) all reach, or come within a few hundredths of, both ends of the valid range, and Late RGB-NIR ($[0.04, 4.96]$) and Early RGB-Only (max 5.00) do so as well despite neither being flagged among the outlier-prone combinations above. This is consistent with the scattered, near-black single-chip outliers described for the CHM-containing combinations qualitatively, but shows that isolated saturated predictions are not exclusively confined to those combinations. All Channels is the only combination under either fusion strategy whose pooled range does not touch either bound (Early: $[0.02, 3.95]$; Late: $[0.28, 3.50]$), reinforcing that it is also the most stable model at the level of individual chip predictions, not only in the spatially coherent appearance of its full-flight map.

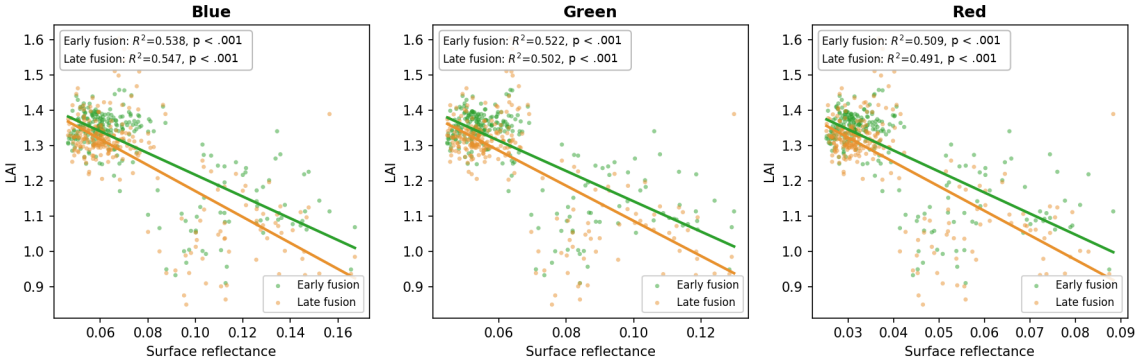


Fig. 5.6: Relationship between UAS-derived LAI and Landsat 8 surface reflectance for the RGB channel-combination model

5.2 Cross-Scale Validation of UAS-Derived LAI Against Landsat 8 Surface Reflectance and Vegetation Indices

This cross-scale comparison was carried out for two of the eight trained channel combinations, RGB and RGB-NIR, as an initial test of the aggregation-and-validation approach, aggregating the RGB (Early and Late Fusion) chip-level predictions from Section 5.1 up to the 30 m Landsat 8 grid and regressing the aggregated LAI against co-registered Landsat 8 surface reflectance and vegetation indices over the same mapped block. For the RGB model (Figure 5.6), aggregated LAI decreases with increasing Blue, Green, and Red surface reflectance under both fusion strategies, the expected direction, since denser canopy absorbs more visible light, with R^2 between 0.49 and 0.55 (all $p < .001$) and Early and Late Fusion producing nearly parallel fits in every band. Against the vegetation indices (Figure 5.7), aggregated LAI increases with NDVI as expected ($R^2 = 0.54$ Early, 0.51 Late), but decreases with EVI ($R^2 = 0.40$ Early, 0.44 Late), the opposite of the positive relationship normally expected between EVI and vegetation density. The spatial maps (Figure 5.8) make this concrete as the low-NDVI strip along the top edge of the block spatially coincides with the low-LAI strip in both the Early and Late Fusion LAI maps, a genuine cross-sensor consistency check that passes, but the EVI map shows almost the opposite spatial pattern to LAI and NDVI, with its lowest values falling exactly where predicted LAI and NDVI are highest, in the bottom-middle of the block. This EVI relationship is inverted relative

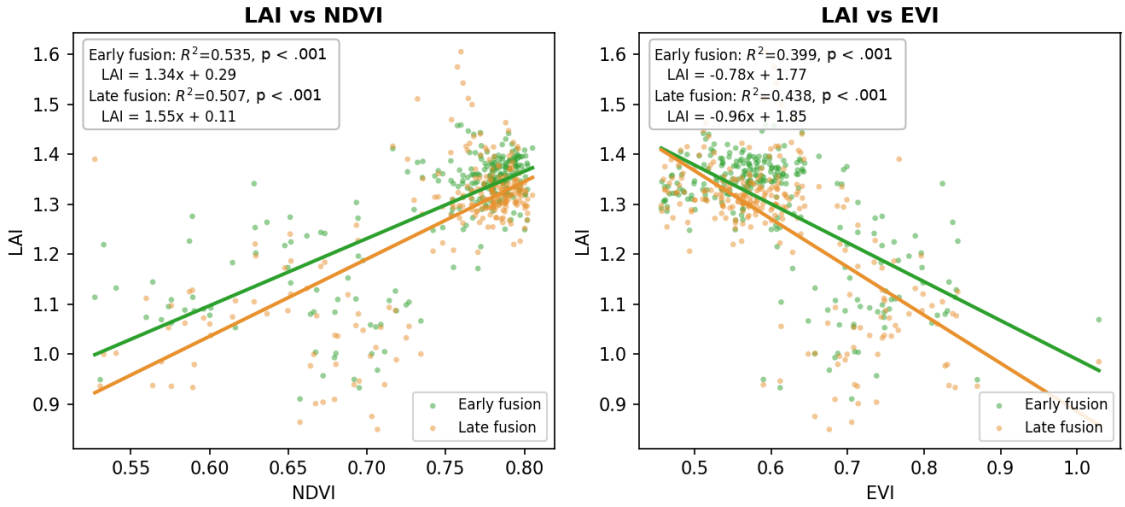


Fig. 5.7: UAS-derived LAI versus Landsat 8 vegetation indices: NDVI and EVI for the RGB model

to physical expectations and is flagged here as an unresolved discrepancy rather than a validated result.

Adding NIR as a fourth input channel (Figure 5.9) preserves the expected negative relationship between aggregated LAI and Blue, Green, and Red reflectance, but the fit is weaker than the RGB-only model for Late Fusion in particular: R^2 falls from 0.55, 0.50, and 0.49 (RGB-only, Blue/Green/Red) to 0.27, 0.32, and 0.32 for RGB-NIR, while Early Fusion's fits stay comparatively stable (0.52, 0.52, 0.50). Against the NIR band itself, the relationship is far weaker still and, for Late Fusion, not statistically significant (Early $R^2 = 0.30$; Late $R^2 = 0.01$, $p = .07$), and where a relationship does hold it is negative rather than the positive one usually expected between NIR reflectance and vegetation density. The same pattern seen for RGB continues here. NDVI is positively related to aggregated LAI ($R^2 = 0.47$ Early, 0.36 Late) while EVI is negatively related ($R^2 = 0.47$ Early, 0.11 Late) (Figure 5.10), confirming that the inverted EVI relationship is not specific to the RGB model. When looking at the maps (Figure 5.11), the Early Fusion RGB-NIR version captures the same low-value edges and higher values in the middle that you see in the RGB map. In contrast, the Late Fusion RGB-NIR map looks much patchier, with scattered pockets of high LAI that don't line up as well with the smoother patterns in the NDVI and

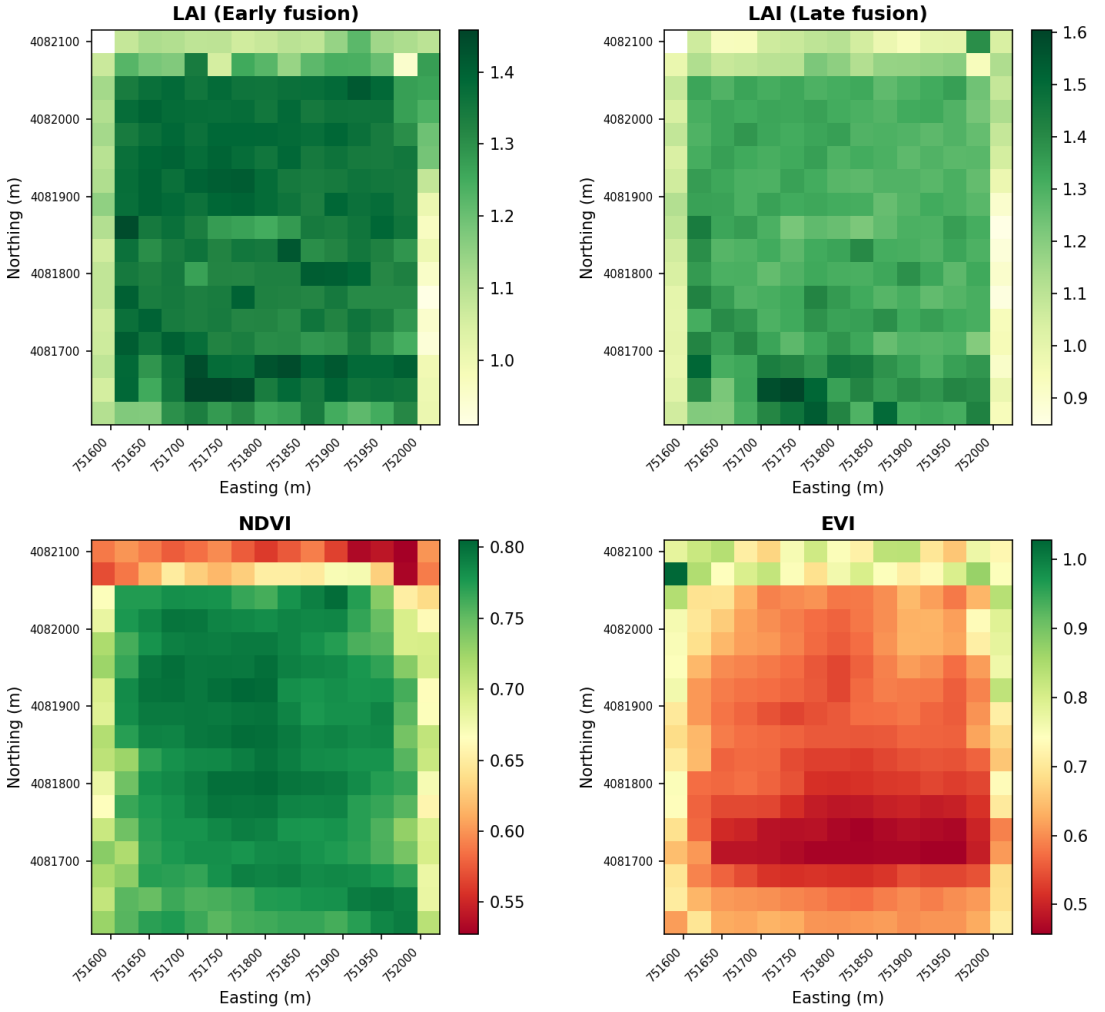


Fig. 5.8: Spatial distribution at 30 m of UAS-derived LAI from the early-fusion and late-fusion models RGB channel-combination model, alongside the co-registered Landsat 8 NDVI and EVI

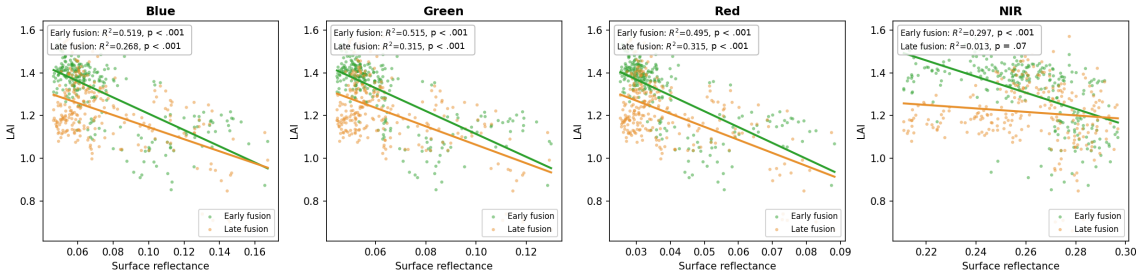


Fig. 5.9: Relationship between UAS-derived LAI and Landsat 8 surface reflectance for the RGB-NIR channel-combination model

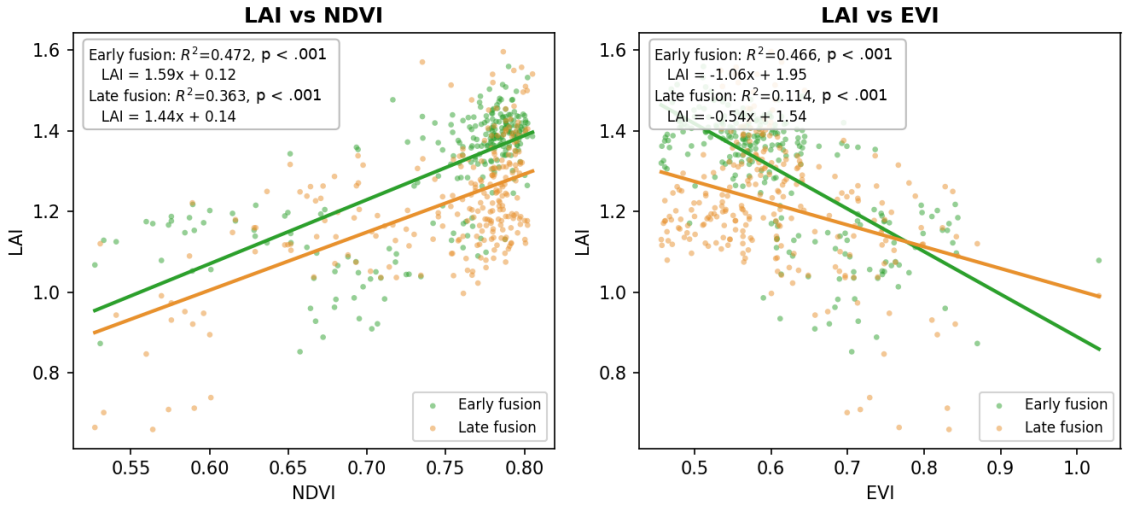


Fig. 5.10: UAS-derived LAI versus Landsat 8 vegetation indices: NDVI and EVI for the RGB-NIR model

EVI layers, especially when compared to the RGB-only Late Fusion map.

Taken together, these results show that the expected relationships between UAS-derived LAI and Landsat 8 surface reflectance and NDVI hold reasonably well at the 30 m scale for both channel combinations tested, that Late Fusion's agreement with Landsat weakens more than Early Fusion's when NIR is added as an input, consistent with the greater sensitivity to individual channels observed for Late Fusion elsewhere in this thesis (Sections 4.1 and 4.2), and that the negative EVI relationship is a consistent, unresolved discrepancy present in both channel combinations rather than a one-off artifact, warranting further investigation before EVI is used to validate UAS-derived LAI at this site.

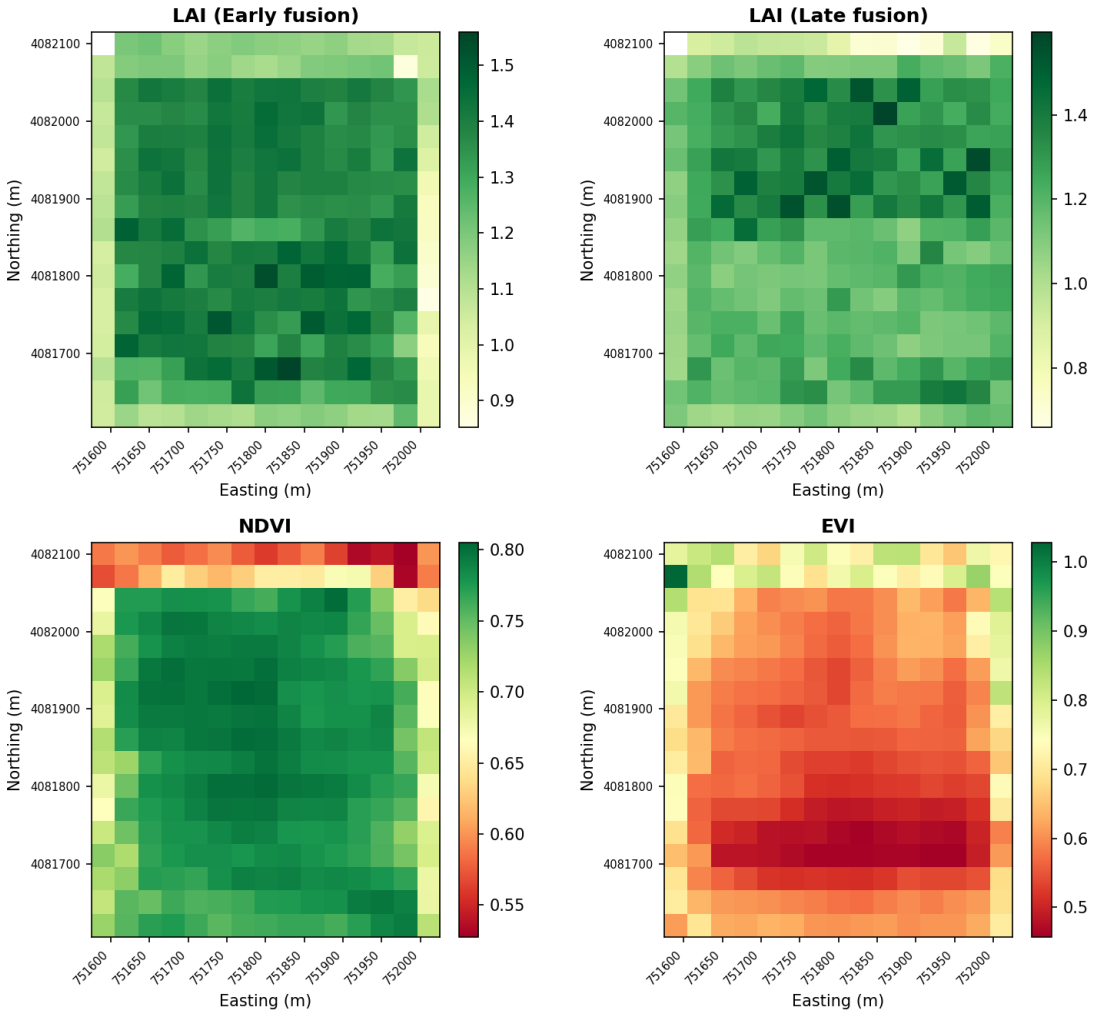


Fig. 5.11: Spatial distribution at 30 m of UAS-derived LAI from the early-fusion and late-fusion models RGB-NIR channel-combination model, alongside the co-registered Landsat 8 NDVI and EVI

CHAPTER 6

CONCLUSION

6.1 Conclusions

This thesis set out to build and evaluate a multi-channel deep learning pipeline for predicting vineyard Leaf Area Index (LAI) from UAS imagery, using ground-truth LAI collected across the Barelli and Ripperdan GRAPEX vineyard sites (Section 3.2) and six co-registered channels, Red, Green, Blue, NIR, DSM, and CHM, derived from each UAS flight (Section 3.3). Every channel was normalized with a Z-score method computed on the training set only (Section 3.4), and all 16 models, every combination of eight channels under both an Early Fusion and a Late Fusion Multi-Channel CNN (Section 3.6), were trained with an identical experimental setup, including a Huber-style LAI loss chosen to reduce sensitivity to noisy field measurements (Section 3.5).

Evaluated on two held-out test flights entirely absent from training and validation, Early Fusion RGB-CHM was the strongest model overall ($R^2 = 0.91$), and Early Fusion occupied six of the top seven positions when all 16 models were ranked by test-set R^2, despite Late Fusion having up to nearly four times as many parameters for the same channel combination (Table 3.2). The training and validation curves confirm that the Late Fusion models that generalized worst, RGB-CHM, RGB Only, and All Channels, were also the ones that trained the longest and ended with the widest gap between training and validation accuracy, consistent with the additional per-modality capacity in Late Fusion being used to fit the training data more closely rather than to generalize better (Sections 4.1 and 4.2). Across almost every metric and both fusion strategies, RGB-DSM was the weakest channel combination, in some cases performing worse than RGB alone. It indicates that DSM by itself does not just fail to support but can actively deteriorate prediction quality, while CHM was consistently the most valuable auxiliary channel, a pattern traced in Section 4.1

to what these two structural channels physically represent: DSM conflates each flight's own terrain elevation with canopy height, while CHM isolates canopy height alone.

Splitting test performance by site revealed a sharper limitation than any channel combination choice could explain. Every model fit the Barelli test flight reasonably well (R^2 up to about 0.6) but collapsed on the Ripperdan test flight (R^2 between 0.04 and 0.16 for every combination and fusion strategy, Section 4.3), including predictions compressing into a narrow band regardless of the true LAI value. Applying all 16 trained models across an entire Ripperdan UAS flight to produce field-scale LAI maps (Section 5.1) showed that the combinations with the weakest held-out chip-level performance, RGB-DSM above all, also produced the most visually extreme, saturated maps at full-flight scale, reinforcing that these were not isolated failures on a few difficult test chips; pooled across the full flight, every model's mean predicted LAI stayed within the physically valid $[0, 5]$ range, but the DSM-containing combinations systematically shifted this mean and its spread upward relative to every other combination (Section 5.1.3). This full-flight view also surfaced two failure modes invisible at the level of individual held-out chips. Under Late Fusion, RGB-DSM and RGB-NIR-DSM both mispredicted a stretch of unpaved road bordering the block as high rather than near-zero LAI, a genuine confusion between canopy and bare ground rather than random noise, and of the eight Early Fusion maps, only RGB-DSM and RGB-DSM-CHM reproduced the sparser-to-denser canopy gradient visible from the top to the bottom of the block in the true-colour imagery, with the remaining six combinations missing this real spatial structure entirely. Finally, when the UAS-based LAI estimates were averaged to match the 30-meter grid of Landsat 8 and compared with satellite data, there was a clear positive link with NDVI and visible-band reflectances. However, in both channel combinations tested, the relationship with EVI was negative and didn't make sense, leaving a question mark instead of confirming the method's accuracy.

These results should also be read in light of the scale of the dataset used to produce them. The 2293 chips extracted across ten UAS flights (Section 3.5) is small by the standards of deep learning more broadly; after setting aside the fixed test split, only 1783 chips

remained for pooled training and validation, and even after 4× label-preserving augmentation (Section 3.5), Late Fusion's largest architecture still has nearly two million trainable parameters (Table 3.2) to fit from that base. This mismatch between how complex the model is and how much data there is helps explain why Late Fusion tended to overfit instead of actually getting better at new examples (Section 4.2). It likely also adds to the bigger gap seen between the Barelli and Ripperdan test results, since a small training set just can't capture all the different canopy conditions found at every site. This limitation is precisely what motivates the transfer-learning direction noted in the literature review, where models pretrained on large natural-image or remote-sensing datasets and fine-tuned on a smaller labeled set have been shown to reach competitive accuracy without training a large architecture from scratch (Chapter 2, [25]), an approach not explored in this thesis but a natural next step given the dataset size available here.

Taken together, this work shows that an Early Fusion Multi-Channel CNN using RGB and CHM information is both the most accurate and the most efficient of the sixteen architectures evaluated for vineyard LAI prediction from UAS imagery, that Late Fusion's additional capacity is not justified by its generalization performance on this dataset, and that model performance depends at least as much on the specific vineyard site and its LAI range as it does on channel combination or fusion strategy.

6.2 Future Work

Several directions follow directly from the limitations identified in this thesis.

- **Training directly on Landsat-resolution imagery.** The cross-scale comparison in Section 5.2 validated UAS-derived LAI by aggregating chip-level predictions from a UAS-trained model up to the Landsat 8 grid after the fact. One straightforward next step would be to train both the Early and Late Fusion models directly on Landsat images, using real LAI measurements if ground truth data is available. Then, their predictions could be compared with those from the UAS-trained models at the same 30-meter resolution. This would help figure out whether the weak, or in EVI's case,

inverted, relationship with Landsat indices is a real limitation of the UAS-trained model, or just a side effect of averaging predictions after training instead of training at the right scale to begin with.

- **Improving generalization to dense-canopy, high-LAI conditions.** Every model in this thesis generalized far worse to the Ripperdan test flight than to Barelli. Because the training data may under-represent the LAI range seen at Ripperdan, and because indirect LAI measurement is itself known to saturate at high LAI (Section 3.2), future work should investigate whether additional training data specifically in the high-LAI range, or a model and loss function designed to be less sensitive to saturation, narrows this gap.

- **Reducing Late Fusion's tendency to overfit.** Since adding more parameters to Late Fusion didn't actually help the models do better on new data, future work could try things like stronger regularization, sharing parameters across parts of the model, or simply using fewer parameters overall. The goal would be to see if a simpler version of Late Fusion can keep its flexibility without being so prone to overfitting.

- **Broader site and flight coverage.** This thesis evaluated held-out generalization using a single test flight per site. Incorporating additional vineyard sites and flights, particularly ones spanning a wider and more continuous range of LAI values, would help distinguish site-specific distinctive features from more general limitations of the modeling approach.

- **Explicitly handling non-vegetated land cover.** The field-scale maps in Section 5.1 showed that RGB-DSM and RGB-NIR-DSM under Late Fusion mispredicted an unpaved access road as high rather than near-zero LAI, since every training chip was drawn from within the vine canopy and none represented bare ground or other non-vegetated cover. Incorporating chips from roads, bare soil, and other non-canopy land cover into training, or masking such areas out of the field-scale maps using an

auxiliary land-cover layer, would help determine whether this is a data-coverage gap or a more fundamental limitation of the DSM-containing combinations.

REFERENCES

[1] J. R. Jensen and K. Lulla, "Introductory digital image processing: A remote sensing perspective," *Geocarto International*, vol. 2, no. 1, pp. 65–65, Mar. 1987.

[2] "Remote Sensing and Image Interpretation, 7th Edition | Wiley," https://www.wiley.com/en-us/Remote+Sensing+and+Image+Interpretation%2C+7th+Edition-p-9781118343289.

[3] K. P. Vadrevu, "Introduction to Remote Sensing (FIFTH EDITION). By J. B.Campbell and R. H.Wynne. Guilford Press, New York, USA, 2011. ISBN 978 160918 176 5. 297 mm × 210 mm, xxxii + 670 pages. Price US$95·00 hardback." *The Photogrammetric Record*, vol. 28, no. 141, pp. 117–118, Mar. 2013.

[4] L. Zhang, L. Zhang, and B. Du, "Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[5] *Space Mission Analysis and Design*.

[6] "Satellite Technology: Principles and Applications, 3rd Edition | Wiley," https://www.wiley.com/en-be/Satellite+Technology%3A+Principles+and+Applications%2C+3rd+Edition-p-9781118636473.

[7] A. D. George and C. M. Wilson, "Onboard Processing With Hybrid and Reconfigurable Computing on Small Satellites," *Proceedings of the IEEE*, vol. 106, no. 3, pp. 458–470, Mar. 2018.

[8] B. Denby and B. Lucia, "Orbital Edge Computing: Nanosatellite Constellations as a New Class of Computer System," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*. Lausanne Switzerland: ACM, Mar. 2020, pp. 939–954.

[9] C. Coopmans, S. Brimhall, R. Goodman, and S. Petruzza, "Scientific timely actionable robotic data operating system (STARDOS): Architecture and progress," in *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping VI*, vol. 11747. Digital Forum: SPIE, Apr. 2021, pp. 63–73.

[10] G. S. CAMPBELL. [Online]. Available: https://metergroup.com/education-guides/the-researchers-complete-guide-to-leaf-area-index-lai/

[11] F. Baret, O. Hagolle, B. Geiger, P. Bicheron, B. Miras, M. Huc, B. Berthelot, F. Niño, M. Weiss, O. Samain, J. L. Roujean, and M. Leroy, "LAI, fAPAR and fCover CYCLOPES global products derived from VEGETATION," *Remote Sensing of Environment*, vol. 110, no. 3, pp. 275–286, Oct. 2007.

[12] R. Gao, A. F. Torres-Rua, M. Aboutalebi, W. A. White, M. Anderson, W. P. Kustas, N. Agam, M. M. Alsina, J. Alfieri, L. Hipps, N. Dokoozlian, H. Nieto, F. Gao, L. G. McKee, J. H. Prueger, L. Sanchez, A. J. Mcelrone, N. Bambach-Ortiz, C. Coopmans, and I. Gowing, "LAI estimation across California vineyards using sUAS multi-seasonal multi-spectral, thermal, and elevation information and machine learning," *Irrigation Science*, vol. 40, no. 4-5, pp. 731–759, Sep. 2022.

[13] T. Dong, J. Liu, B. Qian, L. He, J. Liu, R. Wang, Q. Jing, C. Champagne, H. McNairn, J. Powers, Y. Shi, J. M. Chen, and J. Shang, "Estimating crop biomass using leaf area index derived from Landsat 8 and Sentinel-2 data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 168, pp. 236–250, 2020.

[14] Z. Li, Z. Zhang, Y. Dian, S. Cai, and Z. Chen, "Multi-Scale LAI Estimation Integrating LiDAR Penetration Index and Point Cloud Texture Features," *Forests*, vol. 16, no. 8, p. 1321, Aug. 2025.

[15] W. P. Kustas, J. G. Alfieri, H. Nieto, T. G. Wilson, F. Gao, and M. C. Anderson, "Utility of the two-source energy balance (TSEB) model in vine and interrow flux partitioning over the growing season," *Irrigation Science*, vol. 37, no. 3, pp. 375–388, May 2019.

[16] H. Fang, F. Baret, S. Plummer, and G. Schaepman-Strub, "An Overview of Global Leaf Area Index (LAI): Methods, Products, Validation, and Applications," *Reviews of Geophysics*, vol. 57, no. 3, pp. 739–799, Sep. 2019.

[17] W. A. White, M. M. Alsina, H. Nieto, L. G. McKee, F. Gao, and W. P. Kustas, "Determining a robust indirect measurement of leaf area index in California vineyards for validating remote sensing-based retrievals," *Irrigation Science*, vol. 37, no. 3, pp. 269–280, Dec. 2019.

[18] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[19] L. P. Osco, J. Marcato Junior, A. P. Marques Ramos, L. A. de Castro Jorge, S. N. Fatholahi, J. de Andrade Silva, E. T. Matsubara, H. Pistori, W. N. Gonçalves, and J. Li, "A review on deep learning in UAV remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102456, 2021.

[20] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision Transformers for Remote Sensing Image Classification," *Remote Sensing*, vol. 13, no. 3, p. 516, Feb. 2021.

[21] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers," Nov. 2021.

[22] D. Wang, J. Zhang, B. Du, L. Zhang, and D. Tao, "DCN-T: Dual Context Network with Transformer for Hyperspectral Image Classification," Apr. 2023.

[23] K. Cha, J. Seo, and T. Lee, "A Billion-Scale Foundation Model for Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 19, pp. 16 363–16 378, 2026, preprint arXiv:2304.05215.

[24] J. Li, Y. Cai, Q. Li, M. Kou, and T. Zhang, "A review of remote sensing image segmentation by deep learning methods," *International Journal of Digital Earth*, vol. 17, no. 1, p. 2328827, 2024.

[25] H. Dastour and Q. K. Hassan, "A Comparison of Deep Transfer Learning Methods for Land Use and Land Cover Classification," *Sustainability*, vol. 15, no. 10, p. 7854, May 2023.

[26] S. Toki, C. Coopmans, and A. Rua, "Evaluating Deep Learning Models for Land Cover Classification via Test-Time Augmentation," in *IEEE Aerospace Conference*. IEEE, 2025.

[27] S. Yang, Y. Ge, J. Wang, R. Liu, and L. Fu, "Predicting alfalfa leaf area index by non-linear models and deep learning models," *Frontiers in Plant Science*, vol. 15, p. 1458337, Nov. 2024.

[28] G. Mulero, D. B. Jacques, and D. Helman, "Retrieving leaf area index in wheat fields using unmanned aerial vehicle (UAV)-based LiDAR and hyperspectral imagery," Copernicus Meetings, Tech. Rep. EGU24-4589, Mar. 2024.

[29] J. Arnó, A. Escolà, J. M. Vallès, J. Llorens, R. Sanz, J. Masip, J. Palacín, and J. R. Rosell-Polo, "Leaf area index estimation in vineyards using a ground-based LiDAR scanner," *Precision Agriculture*, vol. 14, no. 3, pp. 290–306, Jun. 2013.

[30] X. Wang, S. Yan, W. Wang, L. Yin, M. Li, Z. Yu, S. Chang, and F. Hou, "Monitoring leaf area index of the sown mixture pasture through UAV multispectral image and texture characteristics," *Computers and Electronics in Agriculture*, vol. 214, p. 108333, Nov. 2023.

[31] H. Qi, B. Zhu, Z. Wu, Y. Liang, J. Li, L. Wang, T. Chen, Y. Lan, and L. Zhang, "Estimation of Peanut Leaf Area Index from Unmanned Aerial Vehicle Multispectral Images," *Sensors*, vol. 20, no. 23, p. 6732, Nov. 2020.

[32] T. Yamaguchi, Y. Tanaka, Y. Imachi, M. Yamashita, and K. Katsura, "Feasibility of Combining Deep Learning and RGB Images Obtained by Unmanned Aerial Vehicle for Leaf Area Index Estimation in Rice," *Remote Sensing*, vol. 13, no. 1, p. 84, Dec. 2020.

[33] G. Togeiro De Alckmin, L. Kooistra, R. Rawnsley, S. De Bruin, and A. Lucieer, "Retrieval of Hyperspectral Information from Multispectral Data for Perennial Ryegrass Biomass Estimation," *Sensors*, vol. 20, no. 24, p. 7192, Dec. 2020.

[34] J. A. Taylor, T. R. Bates, R. Jakubowski, and H. Jones, "Machine-Learning Methods to Identify Key Predictors of Site-Specific Vineyard Yield and Vine Size," *American Journal of Enology and Viticulture*, vol. 74, no. 1, p. 0740013, 2023.

[35] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[36] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, 2022.

[37] X. Zhou, A. Ren, G. Li, J. Li, L. Qiao, J. Yan, and J. Shen, "A systematic review on UAV remote sensing multimodal data fusion strategies with deep learning architectures," in *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*, 2025, pp. 8845–8850.

[38] J. Jiang, X. Feng, F. Liu, Y. Xu, and H. Huang, "Multi-Spectral RGB-NIR Image Classification Using Double-Channel CNN," *IEEE Access*, vol. 7, pp. 20 607–20 613, 2019.

[39] A. Irfan, Y. Li, X. E, and G. Sun, "Land use and land cover classification with deep learning-based fusion of SAR and optical data," *Remote Sensing*, vol. 17, no. 7, p. 1298, 2025.

[40] M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, and F. B. Fritschi, "Soybean yield prediction from UAV using multimodal data fusion and deep learning," *Remote Sensing of Environment*, vol. 237, p. 111599, 2020.

[41] C. Nguyen, V. Sagan, S. Bhadra, and S. Moose, "UAV multisensory data fusion and multi-task deep learning for high-throughput maize phenotyping," *Sensors*, vol. 23, no. 4, p. 1827, 2023.

[42] C. Coopmans, S. Slack, N. Schwemmer, C. Vance, A. J. Beckwith, and D. J. Robinson, "GreatBlue: A 55-Pound Vertical-Takeoff-and-Landing Fixed-Wing sUAS for Science; Systems, Communication, Simulation, Data Processing, Payloads, Package Delivery, and Mission Flight Performance," *Journal of Intelligent & Robotic Systems*, vol. 110, no. 3, p. 98, 2024.

[43] Earth Resources Observation and Science (EROS) Center, "Landsat 8-9 Operational Land Imager/Thermal Infrared Sensor Level-2, Collection 2," 2020, dataset.

[44] M. R. James and S. Robson, "Mitigating systematic error in topographic models derived from UAV and ground-based image networks," *Earth Surface Processes and Landforms*, vol. 39, no. 10, pp. 1413–1420, 2014.