A Novel Framework for Multi-Path Data Fusion in Earth Observation and New Observing Strategies: Applications to Predicting Forest Canopy Height

Mark Moussa[1], James MacKinnon[1], David Harding[1], Matthew Brandt[1]

[1]*NASA Goddard Space Flight Center, Greenbelt, Maryland, USA*

Exponential growth of data from Earth Observation (EO) assets has necessitated the development of sophisticated methods for data interpretation and management. NASA's New Observing Strategy (NOS) approach aims to coordinate operations among complex heterogenous systems of constellations, requiring advanced Artificial Intelligence and Machine Learning (AI/ML) techniques. Despite significant advancements in AI/ML across various domains, the EO and machine learning for satellite (SatML) fields remain fragmented, often relying on adapted techniques rather than domain-specific solutions.

We present a novel end-to-end data fusion framework tailored specifically for EO and SatML, addressing this gap by facilitating rapid development of AI/ML applications. This framework, called, Multimodal Earth Observation Workflow for Machine Learning (MEOW-ML), supports the entire AI/ML lifecycle, from dataset manipulation, to model training, evaluation, and logging, and is designed to expedite the development of next-generation NOS deployments and SOTA in EO.

We apply our framework to predict canopy height model (CHM) derived from lidar data. We integrate multiple data modalities through a hierarchical, multi-path model architecture, effectively identifying and leveraging the unique strengths of each data source to enhance predictive accuracy.

Our experiments demonstrate that the multi-path architecture outperforms traditional single-path models and provides significant advantages in both accuracy and computational efficiency.

1 Introduction

As EO remote sensing capabilities increase, the amount of data available does too. This has led NASA to pursue a NOS approach for coordinating operations between complex sensor webs including combinations of satellite constellations, airborne assets, and ground sensors providing multiple sources of spatial, spectral, temporal, and radiometric data [1]. With such complex systems, comes a need for the intelligent interpretation and management of the data. Thus, AI/ML becomes a valuable and critical tool for enabling next-generation NOS.

AI/ML has advanced rapidly in recent years in many science domains, but the potential for EO and SatML remains relatively untapped. A review by Rolf, et al. [2] posits that SatML should be considered its own distinct modality in order for true state-of-the-art (SOTA) applications for this domain to arise.

To facilitate this, we have developed an end-to-end data fusion framework that handles the entire AI/ML lifecycle, from dataset manipulation, to training, testing, and storing the models and their metrics for evaluation. Our framework, MEOW-ML, is specifically catered to EO and SatML, allowing for the abstraction of Machine Learning Operations (MLOps), and enabling the rapid development of AI/ML applications in this domain.

As a proof of concept for our approach and design, we apply the framework to a forest productivity and degradation use case. For model truth, we use canopy height models (CHM) and canopy height change (CHC) derived from lidar as proxies for productivity and degradation. We use multiple modalities of data for training, with a novel hierarchical, multi-path model architecture that extracts salient features from each data source and merges them automatically in an intelligent way. This allows us to best leverage each data modality's strengths, which provides a more complete picture than one modality, or multiple modalities with a less focused architecture.

2 Methods

2.1 New Observing Strategy

In response to the 2017 National Academy Earth Science Decadal Survey [3] mission design needs, the NASA Earth Science Technology Office (ESTO) has

created the concept of NOS. This describes a variety of potential systems with multiple collaborative instruments, both in space and on the ground, enabling a more dynamic picture of the Earth biological system. Our goal with MEOW-ML is to enable the development of powerful NOS configurations which can further our understanding and modeling accuracy of vegetation function and its relationship to climate change. Our framework can efficiently combine heterogeneous datasets and increase the return-on-investment of existing and future observing systems while connecting global vegetation patterns to underlying physical drivers and processes. By comparing model performance on various combinations of dataset inputs at a variety of spatial, temporal, and spectral resolutions, we can accurately predict the most efficient types of instruments to deploy in an NOS for a given task.

2.2 Dataset

Our dataset has been assembled for 21 NSF National Ecological Observatory Network (NEON) [4] forest sites distributed across the United States, where NEON acquires multi-year airborne remote sensing data including lidar digital terrain and surface elevation models and visible-near infrared-shortwave infrared hyperspectral imaging gridded at 1m resolution [5]. The airborne data at each site typically cover approximately $12km^2$ square areas and are distributed by NEON in $1km^2$ square tiles. To represent forest attributes for training inputs, we have derived forest structure and texture parameters from the lidar elevation models and composition and function parameters from the hyperspectral data. Pre-processing of the spectra includes smoothing, to reduce band-to-band noise, followed by calculation of the 1^{st} derivative, to remove effects on spectra amplitude unrelated to composition and function. A variety of spectral indices were also computed. Additional data sets used for the ML training are gridded soil parameters and climate and drought time series. The data set parameters and their sources are identified in Mackinnon et al., 2023 [6].

2.3 Data Fusion

Data fusion technologies are pivotal in advancing machine learning models for predicting canopy height and height change. We employ a multi-path data fusion methodology, or hierarchical neural network architecture, to process these disparate data sources. Each type of data is fed into separate, specialized pathways before being merged, allowing us to leverage their unique characteristics and extract salient features more effectively. The gridded hyperspectral data, vegetation indices, lidar structure information and soil parameters are each processed through dedicated input layers, followed by training layers optimized for the specific nature of each data type. The features extracted from these individual pathways are then concatenated at the final stage, facilitating the fusion of rich, complementary features from various sources. This integrated approach allows the final layers of the model to make more informed and precise predictions regarding canopy height and height change.

This multi-path fusion technique offers significant advantages over simpler approaches where all data types are processed through a single input layer and identical subsequent layers. When disparate data types are processed together without distinction, the model may fail to capture the unique properties and interactions specific to each data source, leading to suboptimal feature extraction and reduced model performance. In contrast, our multi-path architecture ensures that each data type is processed in a manner that maximizes its individual contributions to the final prediction. Consequently, this approach yields more accurate and reliable predictions of canopy height and height change, which are crucial for monitoring and managing forest ecosystems and understanding environmental dynamics.

2.4 Training Framework

Finding the most optimal configuration of model architecture, hyperparameters, and features requires extensive experimentation, which can be both time-consuming and complex. Evaluating and interpreting model performance and results often becomes overwhelming and hard to track. Given the scarcity of machine learning frameworks tailored for satellite data, especially hyperspectral images, we developed a versatile training framework to address this gap. Our framework, called MEOW-ML, allows users to define their data generator and model architecture configurations, including (but not limited to) hyperparameters and layers, through a simple human readable config file. This config file is extensible, allowing the user to easily add hyperparameters and data manipulation tactics specific to their use case. Built on top of robust open-source tools like Keras, TensorFlow, and MLFlow, our solution is highly customizable. Users can easily extend the framework to create their own training pipelines with different libraries and backends (e.g., PyTorch, JAX), as well as implement their own custom data generators.

This framework not only streamlines the setup and execution of machine learning experiments but also enhances reproducibility and efficiency in model ar-

chitecture development and data processing. By leveraging the flexibility of MEOW-ML, users can quickly iterate over various model and data setups, significantly accelerating the discovery of optimal solutions for their specific satellite data applications.

We will submit MEOW-ML to the NASA open-source approval process, making it available to the community at large. By open-sourcing, end users will be able to leverage existing interfaces by utilizing pre-built components, develop custom interfaces to easily create and integrate new data generators, training pipelines, and model architectures, as well as leverage components created by other users within the open-source community for seamless data generation, model training, and evaluation, all by modifying a configuration file, thereby accelerating the experimental process.

MEOW-ML is designed to be highly modular, allowing researchers to adapt it to their specific needs. For example, users can create a custom data generator for a new type of satellite imagery or develop a specialized training interface to optimize model performance for a unique dataset. Figure 1 shows our example use case and where MEOW-ML specifically helps speed up development.

3 Experimental Setup

In this section, we detail the experimental setup used to evaluate the effectiveness of our machine learning framework for predicting canopy height and height change. Our experiments involved testing various model architectures and applying different data manipulation techniques to optimize performance and accuracy. For the training results presented here, we use 2015 data from the NEON Mountain Lake Biological Station (MLBS) in Virginia, United States applied to prediction of canopy height. To efficiently run and evaluate a large number of modeling cases for comparison purposes, 10 1km^2 cells from the 131 at MLBS were randomly selected.

3.1 Data Manipulation Techniques

During our experimental studies, we further modified the dataset in various ways to increase salience of features and model performance.

3.1.1 Masking

Use of high quality training data relevant to forest productivity and degradation was ensured through the creation of a mask that removes pixels where data is missing or is unvegetated based on low greeness (Normalized Difference Vegetation Index less than 0.5) or where canopy height is less than 0.5m. The masking parameters are easily configurable and can be changed depending on the end user's preference.

3.2 Model Architecture Techniques

The model architecture was a key point of innovation. We started with simple model architectures, and increased in complexity, to judge how well the performance of the model increased accordingly. First, we used Extreme Gradient Boosting (XGBoost) [7], a decision-tree based machine learning framework, to test its efficacy on our dataset and provide a baseline performance metric. Then, we moved to a simple single-path multi-layer perceptron (MLP), concatenating all heterogeneous data types and then running them through the same layers. Then, we experimented with a single-path convolutional neural network (CNN), to gain more insight and localized feature extraction over the spectra. Then, we trained an autoencoder on our high-dimensionality data modalities, namely our hyperspectral data, to reduce the number of features in latent space. We then put the trained encoder into our model architecture, allowing it to increase the signal-to-noise ratio for our input data. Finally, we experimented with various multi-path hierarchical architectures. Figure 2 expands on the hierarchical architecture experiments that were run.

4 Results

In addition to the qualitative assessment, we present detailed quantitative metrics for each model architecture. Table 1 shows the Mean Absolute Error (MAE), and R^2 score for the control XGBoost, single-path MLP, single-path CNN, and multi-path CNN models. The model architectures are listed in Figure 2.

5 Discussion

Our results show that the multi-path model architecture performed better overall, giving slightly more accuracy than the control XGBoost, single-path MLP, and single-path CNN. Specifically, the multi-path CNN, shown in *Case 3* in Figure 2 consistently yielded the best accuracy and generalized the best to unseen tiles in the test set. Interestingly, this model architecture also converged the fastest, requiring less passes over the data (i.e., epochs).

The superior performance of the multi-path CNN (*Case 3*) can be attributed to the multi-path architecture's ability to effectively extract the salient features

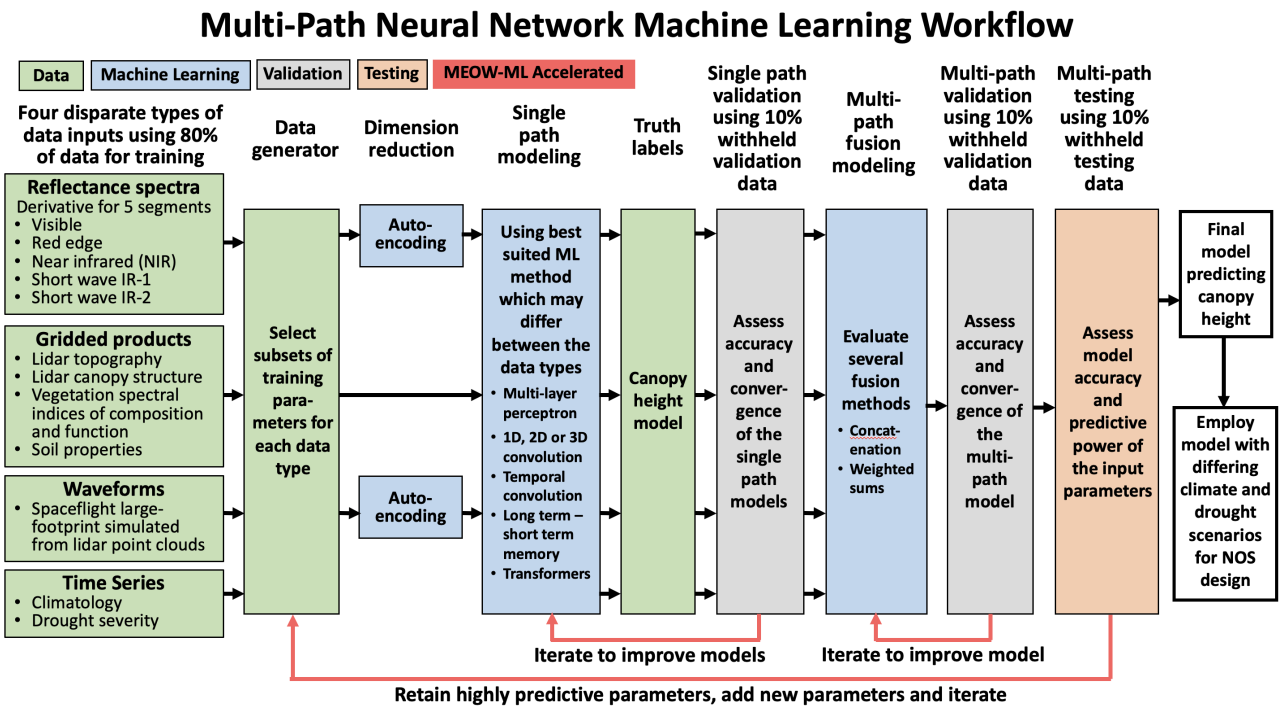

Figure 1: Our use case workflow, from data processing to model development to final model use. The arrows denoting iterative loops (colored in red) show the areas in development which MEOW-ML significantly speeds up.

Model	MAE (meters)	R^2
XGBoost (Case 0)	4.48	−0.36
SP MLP (Case 1a)	6.34	−0.55
MP MLP (Case 1b)	3.92	0.06
SP CNN (Case 2a)	4.52	−0.10
SP E+MLP (Case 2b)	4.10	0.01
MP E+CNN (Case 2c)	4.13	−0.01
MP CNN (Case 3)	3.36	0.26

Table 1: Performance Metrics for Different Model Architectures. All Model experiments can be cross-referenced for further details in Figure 2. SP = Single-path, MP = Multi-path, E = Encoder, MLP = Multi-layer Perceptron, CNN = Convolutional Neural Network.

of the different modalities of input data – hyperspectral data, vegetation indices, and lidar-derived information. By processing these data types through separate paths, the model can extract and integrate relevant features more efficiently than a single-path architecture. This specialized processing allows the model to capture complex relationships within and across each data type, leading to more accurate predictions. The results for our use case show that there seem to be relatively little predictive power between spectra and CHM and indicate that when integrating vegetation indices and lidar-derived data sources, the model performance significantly benefits.

Importantly, the speed and efficiency of the testing and iteration process of the model architecture design and the data pre-processing was significantly enhanced by MEOW-ML. We were able to rapidly iterate over numerous hyperparameters, try different model architecture designs, and toggle data pre-processing techinques, all from a configuration file. Therefore, we were able to arrive at the conclusions we did at a greater pace.

We believe that MEOW-ML has the potential to be generalized to various Earth science applications utilizing satellite data and similar modalities. Given the ability of hierarchical architectures to extract salient features from multiple modalities, we expect that this approach can be extended to many other applications of satellite data or similar modalities in the Earth science domain. For example, the framework can be generalized to be used in cases such as (but not limited to): soil moisture estimation by combining microwave and optical data sources; land cover classification through the integration of multispectral and topographic data; natural disaster detection, such as floods or wildfires, by combining optical and thermal data; coastal monitoring through the integration of sea surface temperature and lidar measurements; atmospheric compo-

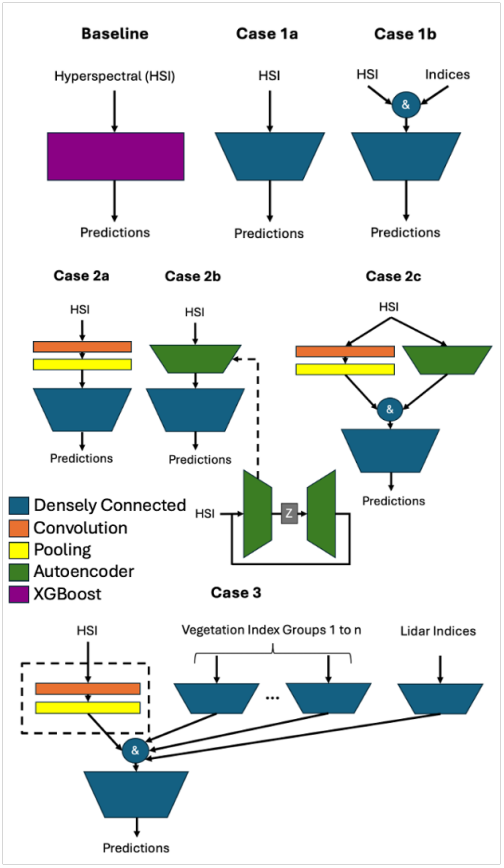

Figure 2: We leveraged MEOW-ML to efficiently run many model configurations. It includes XGBoost (Baseline), single-path MLP (Cases 1a and 1b) and CNN (Cases 2a) models. Additionally, we tested encoded inputs (Case 2b) and combined encoded and CNN inputs (Case 2c). Finally, we tested the fully multi-path model (Case 3). The multi-path models process different data types through specialized pathways before merging features. Architectures with encoders incorporate pre-trained modules for dimensionality reduction, enhancing feature extraction and predictive accuracy.

sition monitoring using satellite and ground-based sensors; and tracking snow and ice cover by fusing optical and SAR data.

The model performance in this case study will be improved upon by further data augmentation and manipulation, and model architecture design consideration.

6 Conclusion

In this paper, we introduced MEOW-ML, a novel AI/ML learning framework built upon a suite of open-source technology and custom components and geared towards EO and SatML. MEOW-ML handles the full lifecycle of ML, from pre-processing of data, to training the model, to logging, testing, and evaluating the model. We used this framework to develop a multi-path, hierarchical neural network, to predict forest productivity and degradation by using canopy height as proxy. We achieved promising results. Our experiments demonstrate that MEOW-ML rapidly accelerated the prototyping phase, allowing us to greatly increase the types of models under investigation. This work enables novel capabilities for NASA's NOS, as well as provides a framework for rapid testing within the EO and SatML scientific community, so new SOTA learning techniques can further enhance this domain.

Acknowledgements

We would like to thank the NASA Earth Science and Technology Office for funding this work, proposal #NNH21ZDA001N-AIST, under the Advanced Information Systems Technology (AIST) program.

Resources supporting data preparation and manipulation were provided by the NASA High-End Computing (HEC) Program through the NASA Computational Information Science and Technology Office at Goddard Space Flight Center.

References

1. Le Moigne, J., Little, M. M., Cole, M. & Ellis, J. New Observing Strategies (NOS) for Future NASA Earth Science Missions. **2019,** IN23C–19. https://ui.adsabs.harvard.edu/abs/2019AGUFMIN23C..19L (2024) (Dec. 2019).
2. Rolf, E., Klemmer, K., Robinson, C. & Kerner, H. *Mission Critical – Satellite Data is a Distinct Modality in Machine Learning* en. Feb. 2024. http://arxiv.org/abs/2402.01444 (2024).
3. *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space* ISBN: 978-0-309-46757-5. https://www.nap.edu/catalog/24938 (2024) (National Academies Press, Washington, D.C., Dec. 2018).
4. *Home | NSF NEON | Open Data to Understand our Ecosystems* https://www.neonscience.org/ (2024).
5. Musinsky, J. *et al.* Spanning scales: The airborne spatial and temporal sampling design of the National Ecological Observatory Network. en. *Methods in Ecology and Evolution* **13,** 1866–1884. ISSN: 2041-210X. https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13942 (2024) (2022).
6. MacKinnon, J. *et al. Multi-Path Fusion: A Hierarchical Machine Learning Approach for Combining Diverse Data Sets for a Forest Monitoring New Observing System* en. in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium* (IEEE, Pasadena, CA, USA, July 2023), 1708–1711. ISBN: 9798350320107. https://ieeexplore.ieee.org/document/10282678/ (2024).
7. Chen, T. & Guestrin, C. *XGBoost: A Scalable Tree Boosting System* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, Aug. 2016). http://dx.doi.org/10.1145/2939672.2939785.