SDHCN: Fire-Ready Forests Data Challenge Report

Shuyu Wang
University of California, San Diego
La Jolla, California
shw043@ucsd.edu

Caroline Zhang
University of California, San Diego
La Jolla, California
caz020@ucsd.edu

Yongyi Jiang
University of California, San Diego
La Jolla, California
yoj001@ucsd.edu

1 Introduction

Efficient and scalable methods for assessing forest composition are critical for ecological monitoring and wildfire modeling. Traditional field-based data collection provides high-quality information but is often limited by time, cost, and spatial coverage. Remote sensing technologies offer an alternative by capturing detailed structural information at fine resolution. However, predicting ecological attributes such as plant functional type (PFT), genus, and species directly from TLS-derived data remains a significant challenge.

This program aims to develop a predictive framework for estimating tree-level ecological classifications—specifically PFT, genus, and species—using structural attributes derived from TLS scans. The objective is to enrich TLS-derived tree lists with biologically meaningful labels that can support broader ecosystem modeling tasks, like modeling wildfires.

We trained a multi-stage classification model using Random Forests, based on a combined dataset of field observations and FastFuels outputs. Models used tree height, diameter, and site identity as input features. Predictions were evaluated by comparing the distribution of inferred labels to field-based references at the plot level, and final models were applied to the TLS dataset to estimate species composition across sites.

2 Methodology

2.1 Dataset Description

We used two primary datasets for model training: (1) field-collected treelist data and (2) FastFuels (FF) treelist data [1]. The field dataset includes tree-level measurements and species annotations sourced from `01_plot_identification.csv` and `03_tree.csv`, while the FF dataset is derived from TLS-based treelist data in `FF_treelist_all.csv`. To support classification tasks, we incorporated additional taxonomic information from `FIATreeSpeciesCode_pft.csv` [2], which provides standardized mappings from species codes and names to genus and Plant Functional Types (PFTs).

To construct the final dataset, we integrated the field and FastFuels treelist data into a unified structure. The field data comprises in-situ measurements of individual trees—including height, diameter, and species—collected through standardized forest inventory surveys across multiple sites in California. In contrast, the FastFuels data, generated through LiDAR-based structural analysis and modeling, includes estimated tree dimensions and associated species codes. Using the species-to-genus-PFT mapping file, we harmonized both datasets by assigning consistent taxonomic and functional labels. Following data cleaning, filtering, and standardization, the final dataset contained over 523,375

Preprint. Under review.

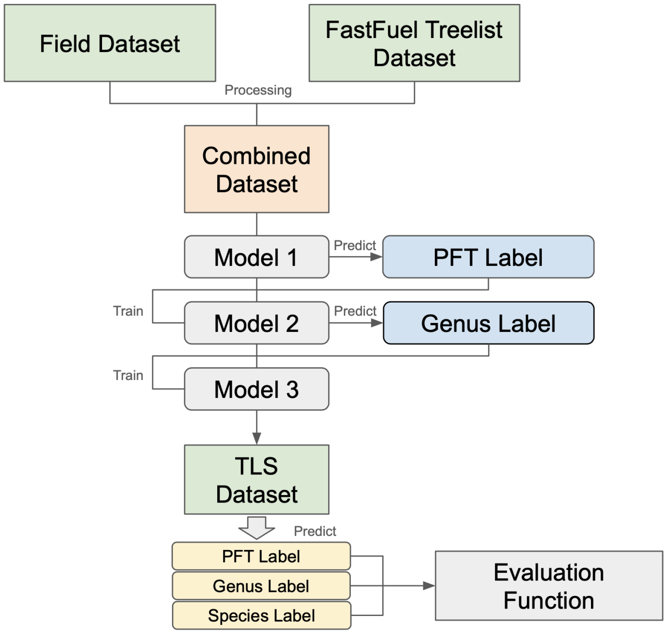

Figure 1: Sample Data Pipline

labeled entries, each annotated with height, diameter, genus, species, PFT classification, and site metadata. Our model pipline is illustrated in Fig 1.

2.2 Data Processing

The data preprocessing pipeline aimed to harmonize and integrate field-collected and FastFuels (FF) data for model training.

For the field data, we began by normalizing scientific names and merging tree records with site metadata using the common identifier, `inventory_id`. Records lacking key attributes such as tree height (`tree_ht`), diameter (`tree_dbh`), or Plant Functional Type (PFT) annotations were removed. We standardized column names to `HT` and `DIA`, extracted genus information from the first token of the scientific name, and converted site names to lowercase for uniformity.

For the FF data, we linked species codes (`SPCD`) with the PFT mapping and filtered out incomplete records. We harmonized column names, extracted genus and full scientific names from the `SCI_NAME` field, and normalized site names in a manner consistent with the field data.

After processing the individual datasets, we concatenated them into a unified table. This final table included the following fields: `HT`, `DIA`, `PFT`, `GENUS`, `species_name`, `site_name`, `inventory_id`, and `source`.

To ensure consistency across the combined dataset and improve the robustness of the model, we applied a series of additional preprocessing steps. Categorical fields, such as `site_name`, `PFT`, `GENUS`, and `species_name`, were normalized by standardizing case (title casing) and trimming any leading or trailing whitespace. Missing values in critical attributes (`HT`, `DIA`, `PFT`, `GENUS`, `species_name`, and `site_name`) were excluded to ensure that all records were complete for training.

2.3 Model Training

We used a hierarchical, multi-stage classification approach based on Random Forests to predict tree-level ecological attributes in three steps: Plant Functional Type (PFT), Genus, and Species. Each stage used the `RandomForestClassifier` implementation from `scikit-learn`, with model hyperparameters set manually based on empirical performance rather than grid search.

Training data consisted of a combined dataset from field observations and FastFuels predictions. Prior to training, text fields were normalized, missing values were removed, and categorical variables—such as PFT, genus, species, and site name—were encoded using `LabelEncoder`. Structural features, specifically tree height (`HT`) and diameter at breast height (`DIA`), were used as primary predictors. Site identity was also included as a categorical feature (`site_code`).

The PFT model was trained first, using `HT`, `DIA`, and `site_code` to predict encoded PFT labels. Its output was then included as an input feature to the genus model, which predicted encoded genus labels using `HT`, `DIA`, `PFT_encoded`, and `site_code`. Lastly, the species model used the full set of predictors—`HT`, `DIA`, `PFT_encoded`, `GENUS_encoded`, and `site_code`—to predict encoded species labels.

Each model was trained with specific hyperparameters: 100–200 trees (`n_estimators`), a maximum tree depth of 10–20, and a minimum split size of 2–5 samples. Cross-validation was used to evaluate model performance, and classification reports were generated to assess precision, recall, and F1-score. Final predictions were applied to the TLS-derived FastFuels dataset to infer ecological composition at the tree level.

3 Result

Our hierarchical model demonstrated strong performance in predicting Plant Functional Types (PFT), but substantially weaker performance on finer-grained classifications such as genus and species. This result held consistently across sites and highlights an important limitation of using structural LiDAR features alone for detailed taxonomic inference. The discrepancy between model performance at different taxonomic resolutions reflects the diminishing signal-to-noise ratio in structural traits when attempting to resolve more specific ecological categories.

3.1 Model Evaluation

To evaluate the performance of the trained models, we applied them to the TLS dataset [3], which, while matching the field dataset in terms of plot locations, lacks ground-truth ecological labels. Therefore, standard accuracy metrics such as F1-scores were not directly applicable.

Instead, we assessed the model outputs by comparing the predicted distributions of Plant Functional Type (PFT), genus, and species in the TLS plots against their corresponding field-derived distributions. Specifically, the model predictions for each tree in a given TLS plot were aggregated to compute the relative frequency of each PFT, genus, and species. These predicted frequency distributions were then compared to the empirical distributions from the matched field plots.

For a quantitative evaluation, we computed summary statistics including the mean and variance of relative abundance for each label across plots. Additionally, the divergence between the TLS-inferred and field-observed distributions was measured using the Kullback-Leibler (KL) divergence, which quantifies the difference between two probability distributions.

KL Divergence Analysis Between TLS Predictions and Field Data

To assess the distributional similarity between TLS-derived predictions and field-observed data, we computed KL divergence at three taxonomic levels: species, genus, and plant functional type (PFT). The figure below presents a boxplot illustrating the range, interquartile spread, and outliers of KL divergence for each level across all plots.

As shown in Fig 2 The results reveal that KL divergence at the **species level** is consistently high, with a wide interquartile range and maximum values exceeding 10. This indicates substantial divergence between TLS-predicted and field-observed species composition, reflecting the model's limitations in resolving fine-grained taxonomic distinctions.

At the **genus level**, the KL divergence is substantially lower, with a median close to zero and modest variability across plots. This suggests that the model is relatively more reliable when aggregating predictions at the genus level, where morphological and structural distinctions captured by TLS are more generalizable.

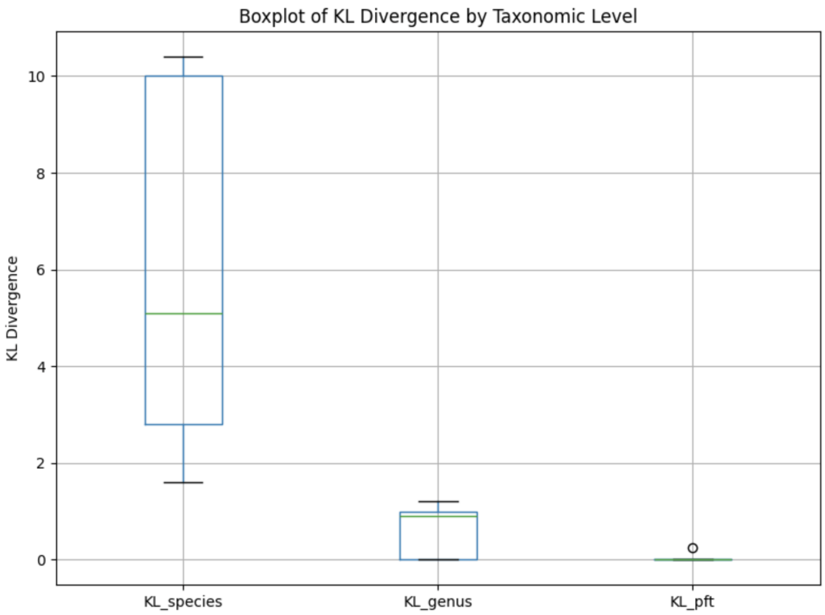

Figure 2: KL Divergence Graph

The **PFT level** shows the lowest KL divergence, with values near zero across nearly all plots and very minimal spread. This indicates strong agreement between TLS predictions and field data at the functional type level, likely due to the broader classification categories and more distinct structural features that TLS can detect at this scale.

Table 1: KL Divergence per Plot Block by Taxonomic Level

Plot Block	KL (Species)	KL (Genus)	KL (PFT)
CATNF_6041_20240814_1	10.42	1.23	0.00
CATNF_6042_20240816_1	2.86	1.00	0.00
CATNF_6043_20240814_1	5.18	0.01	0.00
CATNF_6044_20240817_1	1.64	0.01	0.00
CATNF_6045_20240815_1	10.02	0.87	0.25
... (omit middle rows for brevity)			
CATCU_0120_20241012_1	11.62	7.54	0.76
CATCU_0121_20241012_1	6.63	3.98	0.00
CATCU_0122_20241012_1	17.23	16.82	0.47
CATCU_0123_20241013_1	12.75	16.09	2.08
CATCU_0125_20241013_1	10.61	1.83	0.00

In summary, the model demonstrates high prediction at broader ecological scales (PFT and genus), but species-level classification remains more error-prone.

4 Discussion

4.1 Are the predicted distributions representative of the actual field data?

While predicted PFT distributions from the TLS-derived FastFuels dataset were generally consistent with those observed in the field data, the predicted genus and species distributions showed noticeable differences. This issue is mainly due to variations in the relationship between height and diameter measurements across TLS and field datasets, even within the same plot. These structural inconsistencies affect the model's ability to generalize accurately. Moreover, our evaluation method, which compares distributions at the plot level, may either overestimate or underestimate performance depending on how similar the TLS and field data are in structure. These findings highlight the importance of better alignment techniques or domain adaptation strategies when using field data to validate model predictions.

4.2 Was the field-collected data representative of the surrounding site?

The field-collected data is accurate and appears to be up to date; however, the number of labeled trees per site is relatively small. This limited sampling restricts the ability to fully represent the structural and taxonomic diversity present within each region. Consequently, models trained on this data may struggle to generalize when applied to broader or more complex TLS-derived plots. While the species annotations themselves are reliable, the sparse coverage reduces the field data's effectiveness as a comprehensive reference—particularly for evaluating model performance at the genus and species levels.

4.3 What strategies or techniques could improve the current pipeline?

To address the lack of labeled data and improve model accuracy, we propose adding a semi-supervised learning step to the pipeline. In particular, we suggest using FastFuels plots that match field plots (but do not show their labels) as pseudo-labeled data. By applying our model to these plots and selecting high-confidence predictions, we can build a larger training dataset that better represents the structure of the TLS data. This bootstrapped learning method can help reduce the gap between training and test conditions. Additionally, adding more advanced features—such as canopy shape, crown width, or other LiDAR-based metrics—could help the model distinguish between species that have similar height and diameter.

In addition, we attempted to georeference the plots from the FastFuel dataset to align them with the `plot_blk` identifiers used in the field and TLS datasets. However, the spatial distribution of FastFuel plots did not consistently correspond with the locations in the other datasets. In several cases, plot coordinates differed by 5–6 kilometers from their closest matching site, resulting in significant spatial mismatches and inconsistencies across datasets.

4.4 What new types of data could be included to enhance predictive power?

In the future, combining LiDAR data with other sources could significantly improve prediction quality. Hyperspectral and high-resolution RGB imagery can offer information about leaf color, texture, and other visible traits that are often linked to species identity. Environmental data—such as soil type, moisture, elevation, and disturbance history—can also provide useful context for classification, especially in cases where multiple species share similar structural characteristics. Including these types of data would create a more complete ecological profile and support more accurate predictions at finer taxonomic levels.

4.5 What worked well, and what were the limitations?

What worked well. The integration of field and TLS-derived data was key to improving model performance. This approach expanded spatial and ecological coverage and captured a broader range of tree traits. Classification results were most reliable at higher taxonomic levels. KL divergence analysis showed minimal divergence for plant functional types (PFT) and low divergence at the genus level, indicating strong agreement between TLS predictions and field observations. These results suggest that TLS structural data can effectively support classification at coarse taxonomic scales.

Limitations. Field data, while accurate, was limited in spatial extent and subject to inconsistencies such as protocol variation and missing values. FastFields data, though more extensive, is indirectly derived and may not fully reflect ground conditions due to differences in resolution and processing methods. The TLS feature set is also limited, primarily capturing tree height and diameter. Derived metrics, such as height-to-diameter ratio, offer little additional information. This restricts model performance at finer taxonomic levels, as reflected by consistently high KL divergence at the species level. Further improvement will likely require additional features beyond basic structure.

5 Acknowledgment

We are really grateful to Pedro Ramonetti, University of California, San Diego, for his continuous support, encouragement, and insightful ideas in this class. Many of ideas in this report are proposed and taught by him during the office hour.

References

[1] Anthony Marcozzi et al. “FastFuels: Advancing wildland fire modeling with high-resolution 3D fuel data and data assimilation”. In: *Environmental Modelling and Software* 183 (2025), p. 106214. DOI: `10.1016/j.envsoft.2024.106214`. URL: `https://doi.org/10.1016/j.envsoft.2024.106214`.

[2] U.S. Department of Agriculture, Forest Service, Northern Research Station. *Forest Inventory and Analysis Database*. Tech. rep. Available only on internet: `https://apps.fs.usda.gov/fia/datamart/datamart.html`. St. Paul, MN: U.S. Department of Agriculture, Oct. 2024.

[3] U.S. Geological Survey. *3D Elevation Program Lidar Point Cloud*. Accessed April 13, 2025. 2023. URL: `https://www.usgs.gov/the-national-map-data-delivery`.