remote sensing

MDPI

Article

Multimodal Classification of Urban Trees into Remote-Sensing-Derived Taxonomic Groups Across the United States

Andrew Larkin *, Matthew Bozigar and Perry Hystad

College of Health, Oregon State University, Corvallis, OR 97331, USA; matthew.bozigar@oregonstate.edu (M.B.); perry.hystad@oregonstate.edu (P.H.)
* Correspondence: larkinan@oregonstate.edu

Highlights

What are the main findings?

- Multimodal remote sensing classified individual trees across 36 U.S. cities and 579 taxa.
- Held-out test set accuracy was 0.793, with a macro-F1 of 0.772.

What are the implications of the main findings?

- Multimodal observations can classify individual trees into broader groups defined by similarities in remotely sensed canopy appearance, structure, phenology, and environmental context.
- Performance declined for unseen cities but remained substantially stronger with multi-site training than with models trained on individual municipal inventories.

Abstract

Urban trees provide important ecosystem services; however, municipal inventories used to characterize urban-forest composition remain incomplete and inconsistent among cities. We developed a multimodal deep-learning framework that combined municipal inventories with tree-centered National Agriculture Imagery Program aerial imagery, airborne LiDAR canopy height, Sentinel-2 time series and phenology, Google Satellite Embeddings, and daily weather. Beginning with 5.43 million records from 60 U.S. cities, quality-control and data-availability screening retained 1.85 million records representing 579 taxa across 36 cities. Because many species and genera had overlapping remote-sensing signatures, a supervised discrimination model identified 11 intermediate groups that were subsequently collapsed into six remote-sensing-derived taxonomic groups. On an independent test set of 161,508 records, the final pruned ensemble achieved 0.793 accuracy, 0.772 macro-F1, and 0.792 weighted-F1. Group-specific F1 ranged from 0.705 to 0.867 among trees with LiDAR-derived vegetation height ≥ 2 m. Gate weights emphasized high-resolution NAIP-based inputs, whereas ablation and single-modality analyses showed that Sentinel-2 and weather time series supplied important complementary temporal information. Mean accuracy and macro-F1 declined to 0.685 and 0.478 when target cities were withheld but remained substantially higher than for models trained using individual cities. This framework supports scalable classification of individual urban trees into broader remotely distinguishable groups, not species-level identifications or ecological functional types. Sentinel-2 measurements provide contextual information at 10 m resolution rather than individual-crown observations.

Check for updates

Academic Editors: Yong Ge, Feng Ling, Junfu Fan and Zhoupeng Ren

Received: 11 September 2026
Revised: 15 October 2026
Accepted: 21 October 2026
Published: 28 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: urban trees; tree classification; multimodal remote sensing; deep learning; NAIP imagery; canopy height model; Sentinel-2 phenology; PRISM climate data; satellite embeddings

1. Introduction

Urban trees provide important ecosystem services that can be both benefits and potential harms to surrounding communities. Benefits include heat mitigation [1], carbon storage [2], stormwater interception [3], and opportunities for recreation and psychological restoration [4]. Potential harms include falling limbs, wildfire spread [5], and allergic responses to pollen [6].

Tree species differ substantially in their benefits, risks, and management costs [7]. They also vary in their susceptibility to pests and pathogens and in their resilience to drought and climate change [8]. Optimizing urban tree inventories therefore requires identifying species mixtures that function well together [7]. Similarly, accurately estimating ecosystem services requires accounting for all tree species, not only the dominant ones.

Accurate tree inventories are essential for urban forest management, ecological research, and environmental health assessments. Municipal inventories remain the most common source of information on individual trees, but their completeness, taxonomic resolution, update frequency, and quality vary substantially among jurisdictions [9]. Creating and updating field-based inventories is expensive and labor-intensive, making consistent national coverage difficult to achieve [9].

Identifying urban trees using remotely sensed data offers a potential means of developing standardized, nationwide inventories. However, this approach also presents substantial challenges, particularly in urban environments. One of the greatest barriers to developing a nationwide model is the high diversity and prevalence of introduced tree species. In an analysis of approximately 5.7 million trees across 63 U.S. cities, naturally occurring species accounted for a median of only 45.6% of trees [9]. The dataset contained more than 1000 species representing over 300 tree genera.

Studies combining hyperspectral imagery and Light Detection and Ranging (LiDAR) have successfully classified common tree species within individual cities. For example, Zhang et al. [10] used fused airborne hyperspectral and LiDAR data to classify 15 common urban tree species in Surrey, Canada. Other studies have combined WorldView imagery and LiDAR with convolutional neural networks to classify a limited number of dominant urban tree species [11–13]. Related LiDAR-based approaches have been used to classify common tree species in Chicago, Illinois [14], and Tampa, Florida [15].

High-resolution satellites with frequent revisit times have also been used to classify species based on temporal signatures. Species differ in the timing of leaf emergence, flowering, seasonal productivity, senescence, and leaf retention. Repeated satellite observations can capture these phenological trajectories even when the spatial resolution of an individual image is insufficient to isolate a small tree crown. Jiang et al. [16], for example, combined multiseason Sentinel-2 imagery with airborne LiDAR to classify common urban tree species across Shanghai. Multiseason inputs outperformed single-season observations, and LiDAR-derived structural information improved classification for most species. Nevertheless, like many previous studies, this analysis was restricted to a single metropolitan region and a limited set of common species. At individual-tree scales, Sentinel-2 observations may contain mixed signals from neighboring vegetation and land cover and should therefore be interpreted as contextual temporal inputs within the multimodal framework.

In addition to their limited geographic generalizability, models that focus exclusively on common species have another important limitation: excluding species that occur in the real world but are absent from the training and evaluation datasets can artificially inflate performance estimates. When applied in operational settings, such models will assign less common, unrepresented species to one of the modeled classes. Although each excluded species may account for only a small proportion of errors, the cumulative error associated with dozens of less common species can be substantial and is rarely captured in model evaluations. For example, if a model has 80% accuracy among the top 10 tree species for a city, but only 70% of trees within that city belong to the top 10 species, then when applied to all trees within the city, the model accuracy would be closer to 80% × 0.70 = 56% accuracy. City-level species classification models restricted to only high-prevalence species also highlight a tradeoff between taxonomic specificity and geographic generalizability as dominant species vary greatly across cities [9]. A species classification model generalizable across the United States would require hundreds of species classification labels—many of which share similar remotely observable traits.

An alternative to predicting individual species is to classify broader groups, or supersets, of species. Predicting supersets may provide an effective compromise—a "Goldilocks zone"—that maximizes differentiation among species while satisfying predetermined thresholds for classification performance, geographic extent, and species coverage. Grouping species by genus may initially appear attractive, but it has important limitations. For example, silver maple (*Acer saccharinum*) flowers before leaf emergence, disperses seeds in spring, and is adapted to frequently inundated floodplains, whereas sugar maple (*A. saccharum*) flowers near leaf emergence, disperses seeds in autumn, and is a highly shade-tolerant species associated primarily with well-drained upland forests [17]. Thus, even congeneric species can differ substantially in phenology, environmental tolerances, and ecological function. In contrast, grouping species according to remotely observable traits, such as crown height, leaf-on duration, and crown shape, may maximize class separability using currently available observations while allowing many species to be represented across broad geographic extents.

In this study, we developed and evaluated a national-scale, multimodal deep-learning framework for predicting individual urban trees. We harmonized municipal inventories from 60 U.S. cities and linked them to tree-centered National Agriculture Imagery Program imagery, airborne LiDAR canopy-height data, Sentinel-2 time series and phenological metrics, satellite embeddings, and daily meteorological data. We used a previously published tree-crown detection model [18] to align the remote-sensing datasets with individual crowns, thereby reducing spatial error associated with inaccuracies in tree-inventory coordinates. Rather than attempting to predict hundreds of individual species, we used a supervised discrimination space to organize species into six reported groups characterized by empirically separable remote-sensing features. We then trained an ensemble classifier to predict these groups and evaluated its performance across taxonomic groups, cities, and tree-height classes. This framework addresses the scale, geographic heterogeneity, and large prediction space that have limited the generalizability of previous urban tree-classification models.

2. Materials and Methods

2.1. Study Design and Data Sources

We developed a multimodal deep-learning framework to classify individual urban trees across the continental United States. Municipal tree inventories provided labeled observations that were linked to aerial imagery, airborne LiDAR, Sentinel-2 time series,

satellite embeddings, and meteorological data. A persistent identifier linked each inventory tree to its corresponding inputs throughout preprocessing and modeling.

2.1.1. Tree Inventory

The source dataset contained approximately 5.43 million records from 60 U.S. municipal inventories compiled and harmonized by McCoy et al. [9] (Table S1, Figures S1 and S2). Records generally included geographic coordinates and a taxonomic label, most commonly at the species level. Scientific names were standardized as genus–species binomials when species-level identifications were available and as genus-level labels otherwise. Entries identified only as "unknown," "other," "species," "sp.," or similar non-taxonomic categories were excluded. Honolulu was excluded because Parameter-elevation Regressions on Independent Slopes Model (PRISM) data were unavailable outside of the contiguous United States.

2.1.2. Remote-Sensing Predictor Variables

We compiled remote-sensing and environmental predictors describing individual crowns, their immediate surroundings, seasonal vegetation dynamics, and regional climate (Table 1, Figure S3). Leaf-on, four-band orthographic imagery—red, green, blue, and near-infrared—was obtained from the National Agriculture Imagery Program (NAIP) [19]. For each city, we selected the acquisition closest to 2022, the median municipal-inventory year (Table S2; Figure S3). Native NAIP resolution ranged from 0.08 to 0.6 m.

Table 1. Data Used to Construct the Multimodal Urban Tree Classifier.

Dataset	Description	Spatial Resolution	Temporal Resolution	Source
Tree inventory	Tree coordinates and genus or species identification	Point	snapshot	McCoy et al. (2022) [9]
Aerial imagery	Leaf-on red, green, blue, and near-infrared orthophotography	0.08–0.6 m	snapshot	USDA NAIP
Sentinel-2	Multispectral observations and vegetation indices	10–20 m	5 days	Copernicus Sentinel-2 via GEE
Airborne LiDAR	City-specific point clouds used to construct surface, terrain, and canopy-height products	1 m *	snapshot	USGS 3D Elevation Program
Satellite embedding	Annual 64-dimensional representations of multisensor Earth observations	10 m	Annual	Google Satellite Embedding V1 via GEE
Daily weather	Daily mean temperature, total precipitation, and maximum vapor-pressure deficit	800 m	Daily	PRISM Climate Group

Abbreviations: NAIP—National Agriculture Imagery Program, USDA—United States Department of Agriculture, LiDAR—Light detection and ranging, USGS—United States Geological Survey, GEE—Google Earth Engine, PRISM—Parameter-elevation Regressions on Independent Slopes Model. * LiDAR point cloud resolution varied by city. LiDAR products for all cities were aggregated to 1 m.

Airborne LiDAR point clouds were obtained from the United States Geological Survey (USGS) 3D Elevation Program [20] and used to derive terrain, surface, and canopy-height products at 1 m resolution (Table S3). Disaster-response acquisitions were excluded. Cities without LiDAR acquired after 2012 were also excluded to limit temporal mismatch; this threshold corresponded to a natural break in acquisition years while preserving geographic coverage (Figure S4).

Sentinel-2 observations from 2021–2023 were retrieved through Google Earth Engine [21] for the 10 m grid cell containing each matched crown center. Observations

affected by clouds, shadows, water, or snow were excluded. Annual 64-dimensional Google Satellite Embedding vectors were also extracted for 2021–2023 at 10 m resolution [22]. Daily mean temperature, total precipitation, and maximum vapor-pressure deficit were obtained from PRISM at 800 m resolution for the same period [23].

2.1.3. Tree Crown Identification

Candidate crown centers were detected from four-band NAIP imagery using the Urban Tree Detector convolutional neural network developed by Ventura et al. [18]. Detections with confidence scores below 0.1 were discarded. Each inventory record was matched to the nearest detected crown center within 5 m, giving a 10 m search diameter consistent with the coarsest tree-linked remote-sensing inputs (Figures S5 and S6). Records without a matched crown were excluded.

2.2. NAIP and LiDAR Processing

All spatial inputs were extracted at the matched crown center rather than the original inventory coordinate, reducing sensitivity to positional errors in municipal inventories.

A 64 × 64-pixel NAIP chip was extracted around each crown center. Inputs contained the red, green, blue, and near-infrared bands and a derived normalized difference vegetation index (NDVI) channel. Nonfinite, all-zero, and saturated pixels were identified using a validity mask. During model fitting, each chip was center-cropped to 28 × 28 pixels to emphasize the focal crown and reduce contributions from surrounding buildings, pavement, lawns, and adjacent vegetation (Figure S7).

LiDAR point clouds were used to derive digital surface and terrain models. Canopy height was calculated as their difference, and the resulting products were aggregated to a 1 m grid. Canopy-height model (CHM) chips were extracted around the same crown centers as the NAIP imagery. The CHM input contained canopy height, vegetation-weighted height, and a valid-data mask (Figure S8). Vegetation weighting used spatially aligned NAIP NDVI, with pixels having NDVI below 0.20 assigned a vegetation height of zero.

Structural summaries calculated from the aligned NAIP and CHM chips included valid-data coverage; mean, 95th-percentile, and maximum height; corresponding height summaries among vegetated pixels; the fraction of vegetated pixels exceeding 5 m; and the fraction having both NDVI greater than 0.20 and height greater than 2 m. Summed canopy height was used as a proxy for canopy volume. The difference between the 95th-percentile vegetation height in the chip center and the surrounding ring described the vertical distinctness of the focal crown (Table S5).

To characterize potential mixed-pixel effects at the Sentinel-2 scale, we also calculated the fractions of each corresponding 10 m cell occupied by NAIP-defined vegetation and by vegetation at least 2 m tall according to the aligned CHM.

2.3. Sentinel-2 and PRISM Time Series and Phenology

Sentinel-2 observations were assigned from the 10 m grid cell containing each detected crown center. Surface-reflectance bands and derived vegetation indices, including the normalized difference vegetation index (NDVI), enhanced vegetation index, red-edge chlorophyll index, and leaf-area-index proxies, were aggregated by acquisition date after invalid observations were removed (Table S6). Indices combining 10 and 20 m bands were resampled to a common 10 m grid before extraction.

To regularize the irregular observation sequences, gaps of at least 10 days between valid observations were filled by linear interpolation at 15-day intervals. Each sequence element included indicators identifying observed and interpolated values, interpolation-gap duration and fraction, and sine and cosine encodings of day of year. A data-quality mask

distinguished observed from interpolated measurements and was supplied to the temporal encoder so that the two were not treated as equivalent. The interpolated trajectory was used for input standardization, visualization, and calculation of annual phenological summaries (Figures S9–S11).

Across the 1,851,968 analytical records, interpolated values constituted 13.5% of usable Sentinel-2 time steps. The median record contained 10.0% interpolated observations (interquartile range: 6.3–17.5%). Overall, 15.0% of records contained at least 25% interpolation, whereas 0.13% contained at least 50%. Interpolation gaps had a median duration of 20 days (interquartile range: 15–29 days); their frequency, duration, and geographic variation are summarized in Figures S9–S11.

Daily PRISM variables were aligned with Sentinel-2 sequence dates and supplied to the joint temporal encoder (Table S4). This provided a common temporal index and missingness mask while limiting sequence length.

Ninety-five phenological summaries were calculated from the standardized Sentinel-2 trajectories. These described central tendency, variability, seasonal amplitude, cumulative greenness, and conditions during the early, middle, and late growing seasons (Table S7).

Because a tree crown may occupy only part of a 10 m Sentinel-2 cell, the extracted signal can also represent neighboring vegetation and impervious surfaces. Potential mixed-pixel exposure was characterized using the fractions of each cell classified as vegetation by NAIP and as vegetation at least 2 m tall by the CHM. These variables represent cell-level vegetation context rather than direct estimates of target-crown coverage.

2.4. Satellite Embeddings

For each matched crown center, annual vectors were extracted from the Google Satellite Embedding V1 dataset at 10 m resolution. Each annual observation comprised a unit-normalized, 64-dimensional representation of the temporal and spatial context captured by multisensor Earth observations. Vectors from 2021, 2022, and 2023 were retained separately and concatenated into a 192-element model input, preserving interannual variation rather than averaging across years.

2.5. Inclusion Criteria and Quality Assessment

Several municipal inventories contained many records sharing identical coordinates, suggesting that locations had been assigned to administrative units or otherwise spatially aggregated. Duplicate-coordinate records were removed, and cities in which more than 50% of georeferenced records had duplicated coordinates were excluded (Table S1, Figure S1). This threshold corresponded to a clear natural break in city-level duplicate prevalence (Figure S2).

Within the remaining cities, records were required to have: (1) valid, unique coordinates; (2) a standardized taxonomic label; (3) a detected crown within 5 m; (4) a valid tree-centered NAIP chip with sufficient center vegetation; (5) adequate LiDAR coverage; (6) Sentinel-2 phenology; and (7) a satellite embedding. The 5 m matching radius defined a 10 m search diameter consistent with the spatial resolution of Sentinel-2 and the satellite embeddings. Taxa represented by fewer than 50 eligible records were excluded.

Sequential city-, record-, quality-, and taxon-level screening reduced the 5,427,599 records from 60 municipal inventories to 1,851,968 analytical records representing 579 taxa across 36 cities (Table S1; Figure S1A). Of these, 1,760,578 records (95.1%) had no retained quality warning. The remaining 91,390 records (4.9%) had low LiDAR-derived vegetation height despite otherwise passing the hard exclusion criteria. These records were retained at half weight during training and evaluated as a separate sensitivity stratum.

All continuous structural, phenological, embedding, and meteorological variables were standardized using means and standard deviations estimated exclusively from the training partition.

2.6. Identification of Remote-Sensing-Derived Taxonomic Groups

A discriminator model was trained using standardized taxonomic labels, with genus-level labels retained when species-level identifications were unavailable. Labels required at least 500 training records, resulting in 280 core taxa. Separate convolutional encoders processed NAIP and CHM imagery, while an interaction encoder jointly represented spatially aligned NAIP and CHM patches. Multilayer perceptrons encoded structural metrics, Sentinel-2 phenology, and satellite embeddings. A masked transformer encoded the raw Sentinel-2 sequence together with temporally aligned daily PRISM variables. Softmax gating combined the modality-specific representations, which were projected into a normalized 64-dimensional discrimination space.

Training minimized a combination of supervised contrastive and class-weighted cross-entropy losses. The contrastive temperature was 0.12, and cross-entropy contributed 0.35 of the combined loss. Class weights followed the inverse square root of training frequency and were capped at 3.0. The model was optimized using AdamW, with checkpoint selection based on validation macro-F1 and early stopping (Table S8).

Before model fitting, core-taxon records were assigned to fixed, taxon-stratified training, validation, and locked test partitions. Training records were used to fit the discriminator. Validation embeddings were used for checkpoint selection, taxon-centroid estimation, evaluation of candidate partitions containing 4–12 groups, group-count selection, and 500-replicate bootstrap stabilization. Candidate partitions were compared using weighted silhouette and the mean and minimum distances among group prototypes. Average-linkage clustering of the bootstrap co-assignment matrix produced the 11 groups.

An additional 299 taxa represented by 50–499 records were subsequently assigned to the nearest fixed group prototype. For each rare taxon, its centroid was calculated from an 80% development subset, and the remaining 20% was reserved for testing. Group assignments were frozen before rare-taxon test performance was examined, and no test record contributed to discriminator fitting, centroid estimation, group construction, or taxon assignment.

2.7. Multimodal Classifier Architecture

The group classifier used modality-specific encoders (Figure 1). Five-channel NAIP chips (four reflectance bands and NDVI) and three-channel CHM inputs (height, vegetation-weighted height, and valid-data mask) were processed by convolutional encoders. Structural metrics, Sentinel-2 phenology summaries, and satellite embeddings were processed by multilayer perceptrons with layer normalization, nonlinear activation, and dropout. A masked sequence encoder processed raw Sentinel-2 observations together with temporally aligned daily PRISM variables and observation-quality indicators.

Each branch produced a 128-dimensional representation. These representations were passed to a gating network comprising layer normalization, two fully connected layers, Gaussian error linear unit activation, and dropout. Softmax-normalized gates produced sample-specific modality weights. The weighted branch representations were summed and passed through a classification head that produced probabilities for the 11 intermediate remote-sensing-derived taxonomic groups.

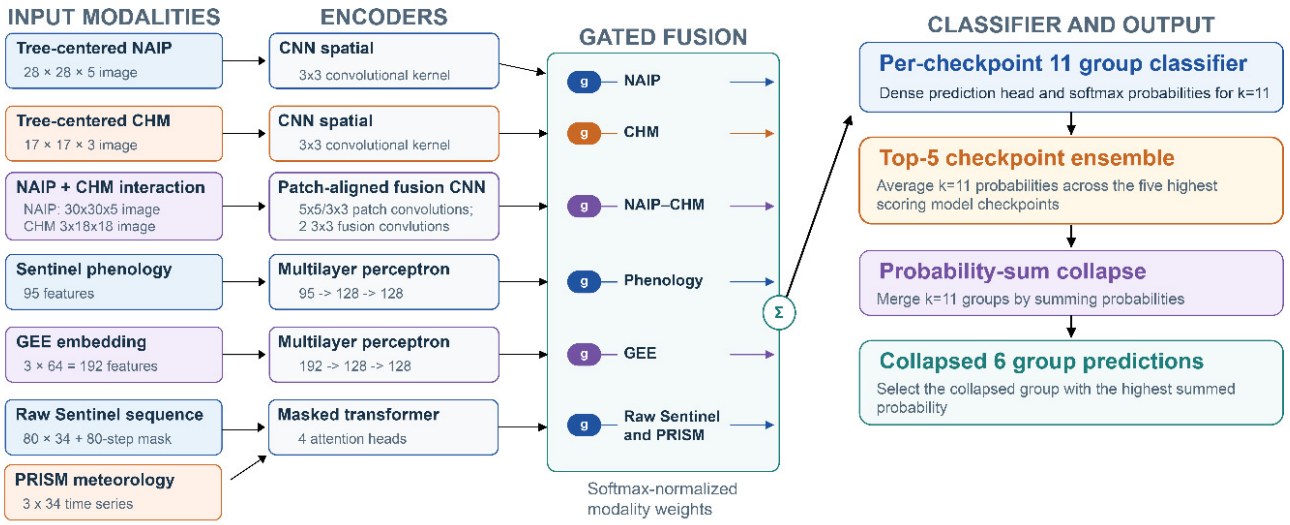


Figure 1. Architecture of the Final Multimodal Urban-Tree Classifier. Six input branches were processed by modality-specific encoders: convolutional networks for tree-centered NAIP and CHM imagery, a patch-aligned convolutional encoder for the NAIP–CHM interaction, multilayer perceptrons for Sentinel-2 phenology and satellite embeddings, and a masked sequence encoder for raw Sentinel-2 observations with temporally aligned PRISM variables. Each encoder produced a 128-dimensional representation. A learned softmax gate assigned sample-specific modality weights, and the weighted representations were summed and classified into 11 intermediate groups. Softmax probabilities were averaged across the five highest-ranked validation checkpoints and then summed according to the fixed six-group collapse to produce the reported predictions. Abbreviations: CHM, canopy height model; NAIP, National Agriculture Imagery Program; PRISM, Parameter-elevation Regressions on Independent Slopes Model.

The intermediate groups were collapsed using a validation-derived hierarchy based on discrimination-space geometry and mutual classification confusion. Candidate partitions containing 5–10 groups were compared using validation accuracy, macro-F1, minimum-group F1, and retained prototype separation, subject to the constraint that no collapsed group contain more than 50% of taxa or records (Table S10). The six-group partition improved separation and accuracy relative to the five-group partition, whereas retaining additional groups reduced accuracy without sufficient gains in separation. Final six-group probabilities were calculated by summing the predicted probabilities of the corresponding intermediate groups.

Candidate branches were evaluated using leave-one-branch-out models trained under otherwise identical conditions (Table S11). Effects were estimated from paired differences between complete- and ablated-model predictions using 2000 record-level Poisson-bootstrap replicates. A branch was pruned only when its exclusion produced a statistically significant performance improvement; nonsignificant differences were treated as insufficient evidence for removal. This procedure removed the tree-centered structural-metrics branch and retained the other six branches.

2.8. Model Training and Evaluation

Fixed, taxon-stratified partitions contained 1,590,460 training, 100,000 validation, and 161,508 test records (Figure S1B). The test set comprised 150,000 locked core-taxon records and 11,508 rare-taxon records reserved before centroid assignment. All group construction, model fitting, modality selection, collapse selection, and checkpoint selection used only training or validation data.

Training images were augmented using synchronized spatial transformations of the NAIP and CHM chips. The model was optimized with AdamW using weighted cross-entropy loss, an initial learning rate of 5×10^{-4}, weight decay of 1×10^{-4}, batch size of 1024, dropout of 0.20, and gradient clipping at 2.0. Square-root inverse-frequency class weights, capped at 3.0, reduced the influence of class imbalance. Sample weights additionally accounted for retained quality warnings and inventory-to-crown matching distance (Table S9).

Models were trained for a maximum of 120 epochs. Checkpoints were ranked using a validation score combining macro-F1 (50%), accuracy (30%), and weighted-F1 (20%). Training stopped when this score failed to improve by at least 1×10^{-4} for 12 consecutive epochs. The five highest-ranked checkpoints were retained. During inference, their 11-class softmax probabilities were averaged and then summed according to the fixed six-group collapse; the group with the greatest probability was selected. The random seed was fixed; thus, this ensemble represents checkpoint averaging within one training run rather than independent stochastic repetitions.

Performance was evaluated on the held-out test set only after all group and model decisions were frozen. Metrics included accuracy, macro-F1, weighted-F1, and group-specific precision, recall, and F1. Results were also stratified by city, taxon provenance, and data-quality criteria. Geographic sensitivity was evaluated using leave-one-city-out models for every city with at least 1000 test records. Single-city models were trained for the seven cities containing at least 10 training records from each intermediate group and evaluated across the eligible target cities. Within-city clustering of classification errors was assessed using Moran's I ($k = 8$) for the city most representative of the median leave-one-city-out accuracy and macro-F1 jointly.

3. Results

3.1. Discrimination Space Group Classification

Figure 2 shows the 11 intermediate groups identified in the discrimination space and their collapse into six reported groups. Group 1 comprised warm-climate evergreen and coastal taxa, including palms, eucalypts, evergreen oaks, *Magnolia grandiflora*, and subtropical ornamentals. Group 2 contained cool-temperate ornamental broadleaf taxa, including maples, cherries, crabapples, pears, and lindens. Group 3 represented temperate deciduous canopy taxa, including ashes, elms, honey locusts, oaks, sweetgum, and silver maple. Group 4 comprised warm- and dry-climate ornamental broadleaf taxa, particularly *Lagerstroemia indica*, *Pistacia chinensis*, warm-climate ashes, and other drought-adapted ornamentals. Group 5 was dominated by *Platanus* and included several riparian broadleaf taxa. Group 6 contained conifers, particularly pines, spruces, junipers, Douglas-fir, and redwood. Assignments for core and rare taxa are provided in Sections S1 and S2, respectively.

The intermediate groups showed substantial but incomplete separation on the independent discriminator test set (Table S12). Nearest-prototype classification among the 11 groups achieved an accuracy of 0.653 and a macro-F1 of 0.649. Collapsing predictions into six groups increased accuracy to 0.763 and macro-F1 to 0.733, with group-specific F1 scores ranging from 0.642 to 0.840. Separation was strongest for the warm-climate evergreen and coastal group (F1 = 0.840). *Pinus ponderosa*, *Platanus acerifolia*, and *Pistacia chinensis* occupied well-separated regions of their respective collapsed groups. By contrast, *Koelreuteria paniculata* and *Populus tremuloides* remained near group boundaries, whereas *Acer macrophyllum* and *Quercus macrocarpa* became substantially less ambiguous after overlapping intermediate groups were combined.

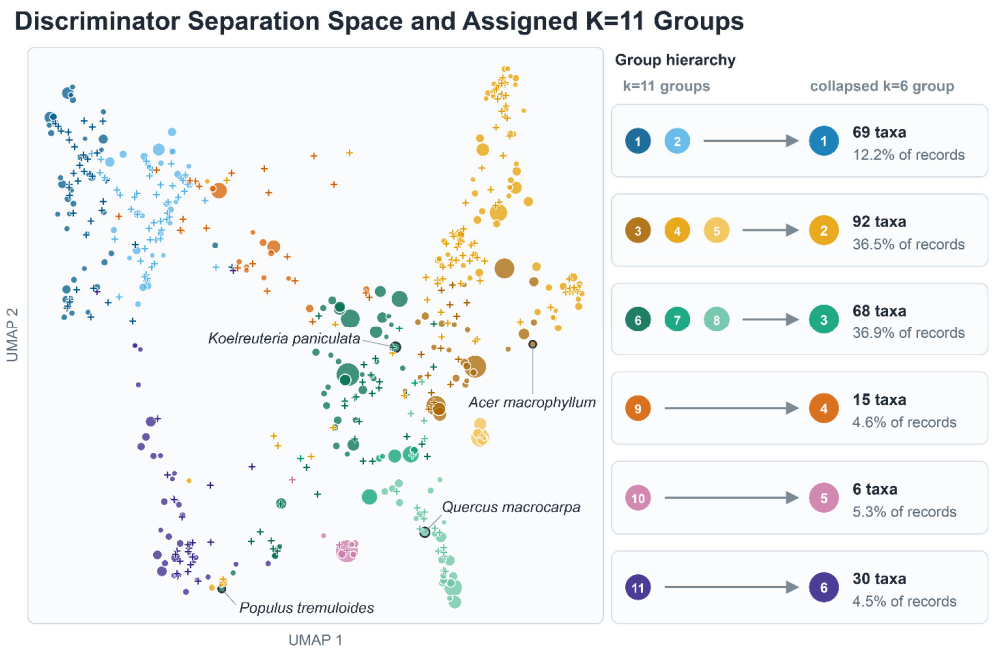


Figure 2. Discriminator Separation Space and Hierarchy of the Remote-Sensing-Derived Taxon Groups. UMAP projection of taxon centroids in the 64-dimensional discriminator space. (**Left**): The 11 intermediate groups used to train the stage-two classifier. (**Right**): Intermediate group collapse into the six reported groups. Identical centroid coordinates are used in both panels. Circles represent the 280 core taxa included in discriminator training, with point area proportional to validation support; crosses represent 299 rare-extension taxa assigned using development-only centroids. The shared hierarchy indicates which intermediate groups contribute to each reported group. UMAP preserves local neighborhood structure, but its axes have no physical units. Abbreviation: UMAP—Uniform Manifold Approximation and Projection.

Group assignments did not consistently follow taxonomic genera, indicating that the learned partitions represented remote-sensing similarity rather than conventional taxonomy. *Acer* taxa occurred across four intermediate groups and two collapsed groups. Most temperate maples, including *A. rubrum*, *A. saccharum*, and *A. platanoides*, occurred in Group 2, whereas *A. negundo* and *A. saccharinum* occurred in Group 3. Most pines were assigned to the conifer-dominated Group 6, but *P. elliottii* occurred in the warm-climate evergreen Group 1 and *P. mugo* in the cool-temperate ornamental Group 2. These within-genus divisions are consistent with differences in remotely sensed canopy appearance, structure, phenology, and environmental context rather than taxonomic relatedness alone.

3.2. Overall Classification Model Performance

Overall performance and sensitivity results for the final pruned model are summarized in Figure 3. On the independent test set, the collapsed six-group predictions achieved an accuracy of 0.793, macro-F1 of 0.772, and weighted-F1 of 0.792. Performance before collapse was lower for the 11 intermediate groups (accuracy = 0.688, macro-F1 = 0.690, and weighted-F1 = 0.688; Table S13), indicating that many errors occurred among intermediate groups subsequently combined within the six collapsed groups.

Performance was highest for records with stronger vegetation-height support and for core taxa represented during discriminator training (Figure 3). Records meeting the 2 m vegetation-height criterion achieved a macro-F1 of 0.777, compared with 0.610 below that threshold. Core taxa achieved a macro-F1 of 0.781, whereas rare taxa assigned after discriminator training achieved 0.624. Macro-F1 was also lower when tall CHM vegetation

occupied less than 25% of the Sentinel-2 cell (0.717 versus 0.794), when NAIP vegetation occupied less than 25% of the cell (0.728 versus 0.786), and when crown-matching distance was at least 2.5 m (0.748 versus 0.793).

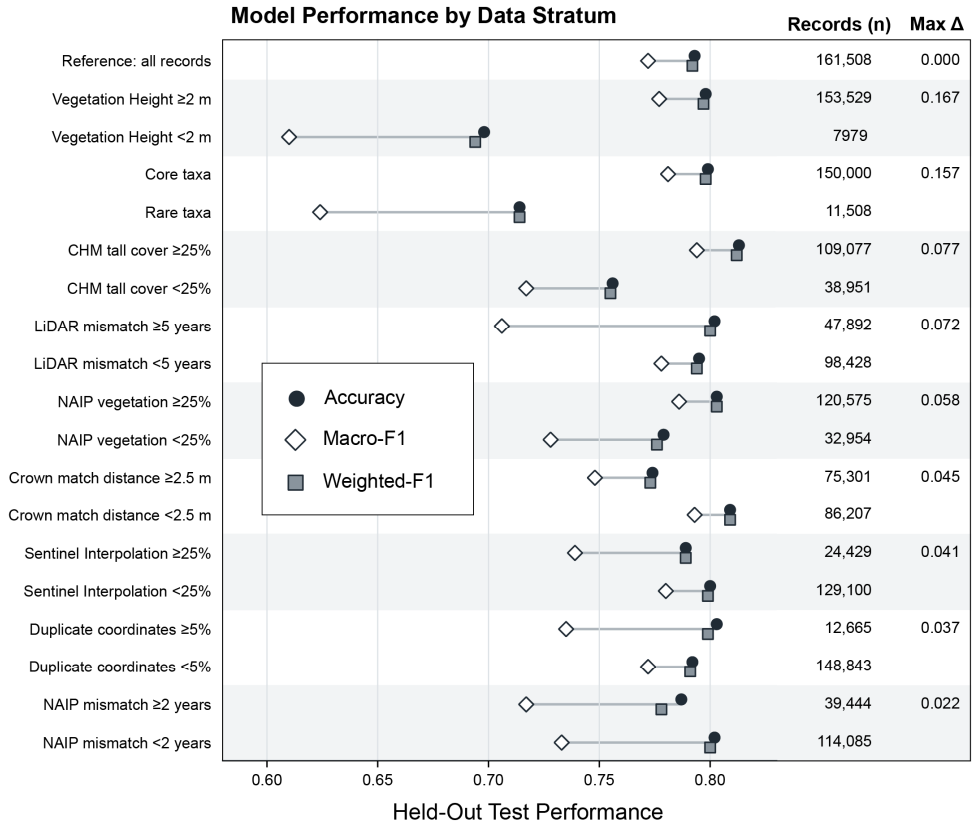


Figure 3. Performance of the Final Pruned Model after Collapsing Predictions into Six Reported Groups, Stratified by Data Quality and Taxon Provenance Criteria. Sensitivity analyses were conducted post hoc on the independent test set without retraining the model. Core taxa had at least 500 records and were included in discriminator training; rare taxa had 50–499 records and were assigned to groups after discriminator training. Abbreviations: NAIP—National Agriculture Imagery Program, LiDAR—Light detection and ranging, CHM—Canopy Height Model. LiDAR acquisition timing, NAIP acquisition timing, and duplicate-coordinate prevalence were largely consistent within cities and were therefore strongly confounded with geographic location. Max Δ is The largest absolute difference between two strata in accuracy, macro-F1, and weighted-F1.

Records with at least 25% interpolated Sentinel-2 time steps achieved an accuracy of 0.789 and macro-F1 of 0.739, compared with 0.800 and 0.780, respectively, among records below that threshold. Thus, extensive interpolation was associated with a modest difference in accuracy but a larger reduction in class-balanced performance. Differences associated with LiDAR and NAIP acquisition years and duplicate-coordinate prevalence were geographically confounded because these properties were largely shared among records from the same city.

3.3. Group-Specific Model Performance

Group-specific performance is shown in Figure 4. This comparison was restricted to records with LiDAR-derived vegetation height ≥ 2 m to reduce potential discrepancies associated with shrubs, juvenile trees, and erroneous crown–inventory matches. F1 was highest for Group 1 (0.867), followed by Groups 2 (0.801), 3 (0.793), and 5 (0.787). Performance was lower for Groups 4 (0.709) and 6 (0.705). Group 1 had higher recall than precision, indicating comparatively frequent overprediction, whereas Group 6 had higher precision

than recall, indicating underprediction. Group-level F1 was moderately but nonsignificantly associated with test support across the six reported groups (Pearson's $r = 0.44$, $p = 0.38$).

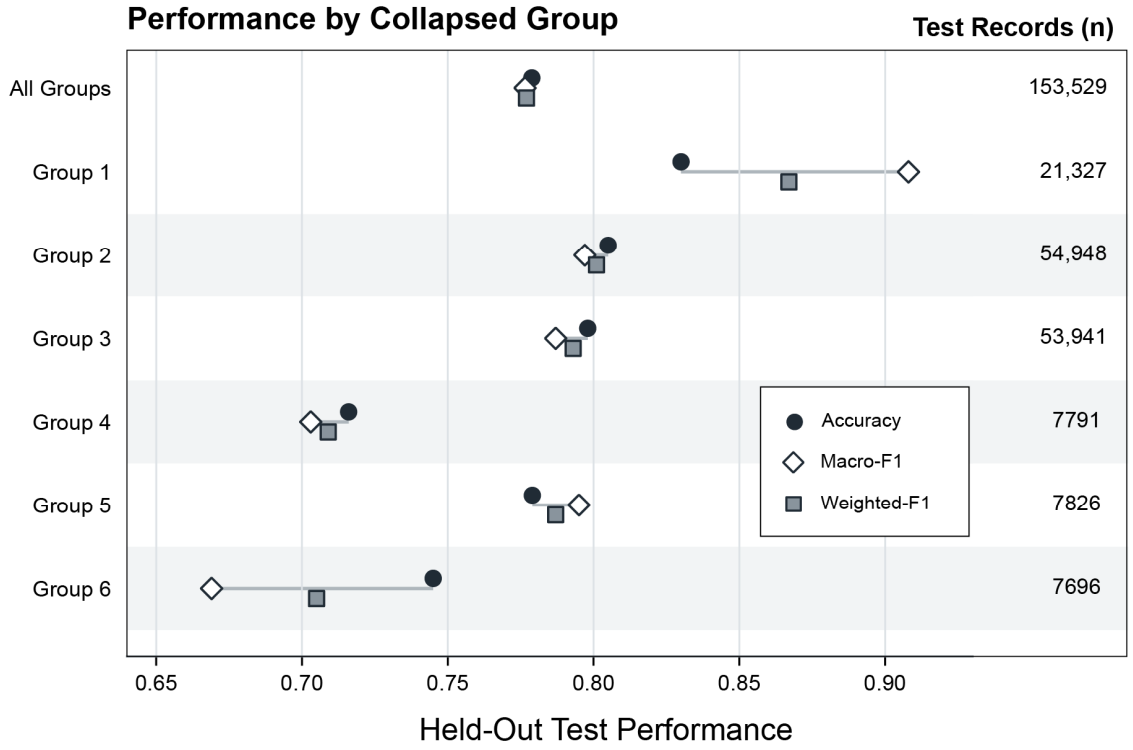


Figure 4. Group-Specific Performance of the Final Pruned Ensemble after Collapsing the 11 Intermediate Groups into Six Reported Groups. Performance was evaluated on held-out test records after excluding records with the vegetation height warning (sample size = 153,529).

3.4. Misclassification and Error

Figure 5 shows the confusion matrix for independent test records with no retained quality warning. Agreement was strongest for Group 1, for which 90.8% of records were correctly classified. Among the remaining groups, diagonal agreement ranged from 66.9% for Group 6 to 79.7% for Group 2. The largest reciprocal confusion occurred between Groups 2 and 3: 14.9% of Group 2 records were predicted as Group 3, and 15.5% of Group 3 records were predicted as Group 2. Other prominent errors included predictions of Group 4 as Group 1 (13.2%), Group 5 as Group 3 (10.4%), and Group 6 as Groups 2 or 3 (13.7% and 9.5%, respectively). Groups 2 and 3 received the largest numbers of incorrect predictions, consistent with their broad, partially overlapping regions in the discrimination space (Figure 2).

Taxon-level error decreased as separation from competing group prototypes increased (Spearman's $\varrho = -0.488$, $p = 3.2 \times 10^{-41}$; Figure 6). Rare taxa exhibited greater variation in error than core taxa, but this difference was concentrated among taxa with unstable or geometrically ambiguous group assignments. Among taxa with at least 80% bootstrap agreement and prototype margins greater than 0.05, mean taxon-level accuracy was nearly identical for core and rare taxa (0.813 and 0.812, respectively), as was record-weighted accuracy (0.824 and 0.815). These results indicate that the lower aggregate performance of rare taxa was concentrated among uncertain group assignments rather than occurring uniformly across taxa excluded from discriminator training.

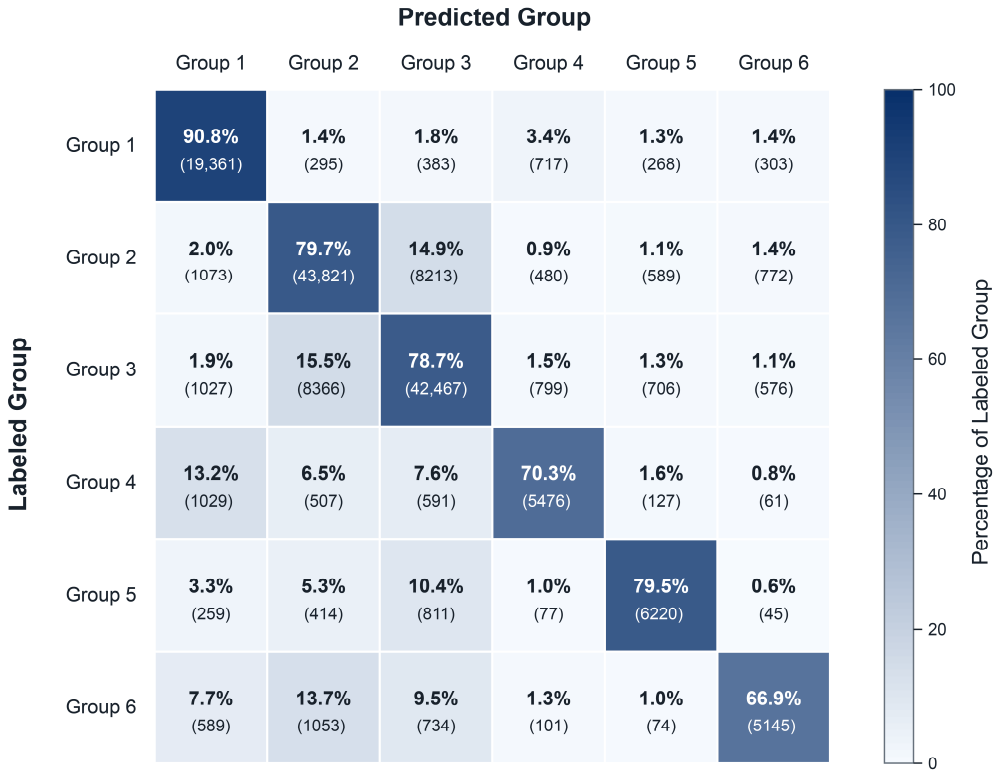


Figure 5. Confusion Matrix for the Classification Model Evaluated on Independent Test Records. Rows represent labeled groups and columns represent predicted groups. Cell shading and percentages are normalized within labeled groups; raw record counts are shown in parentheses.

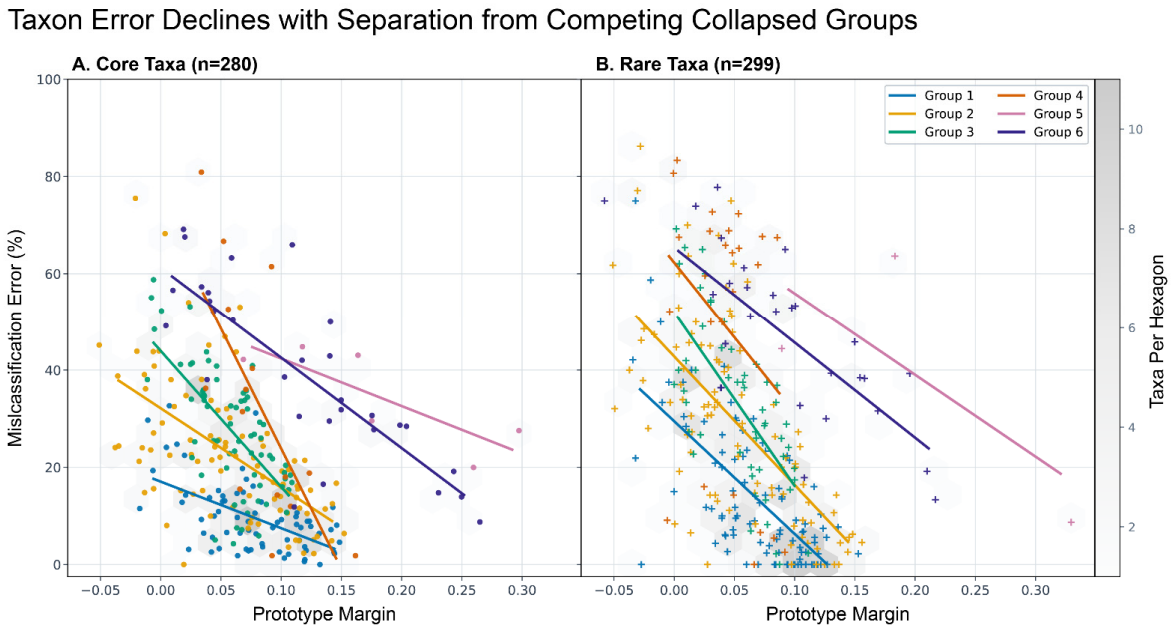


Figure 6. Relationship Between Prototype Margin and Taxon-Level Misclassification Error for the Six Reported Taxonomic Groups. Prototype margin is the difference between a taxon's distance to the nearest competing group prototype and its assigned-group prototype; larger positive values indicate greater separation from competing groups. Core taxa used to train the discriminator are shown as circles (**A**), and rare taxa assigned after discriminator training are shown as crosses (**B**). Prototype margin was negatively associated with error (Spearman $\varrho = -0.488$, $p = 3.2 \times 10^{-41}$).

3.5. Modality Performance and Gate Weights

The final model assigned most gate weight to high-resolution spatial inputs (Figure 7). Tree-centered NAIP received the highest mean weight (0.481), followed by the NAIP–CHM interaction (0.246); together, these branches accounted for 72.7% of the mean gate weight. Mean weights were lower for tree-centered CHM (0.093), satellite embeddings (0.071), Sentinel-2 phenology (0.054), and the raw Sentinel-2–PRISM time series (0.054). Variation in mean gate weights was greater among cities than between vegetation-height or taxon-provenance strata, indicating substantial geographic heterogeneity in modality use.

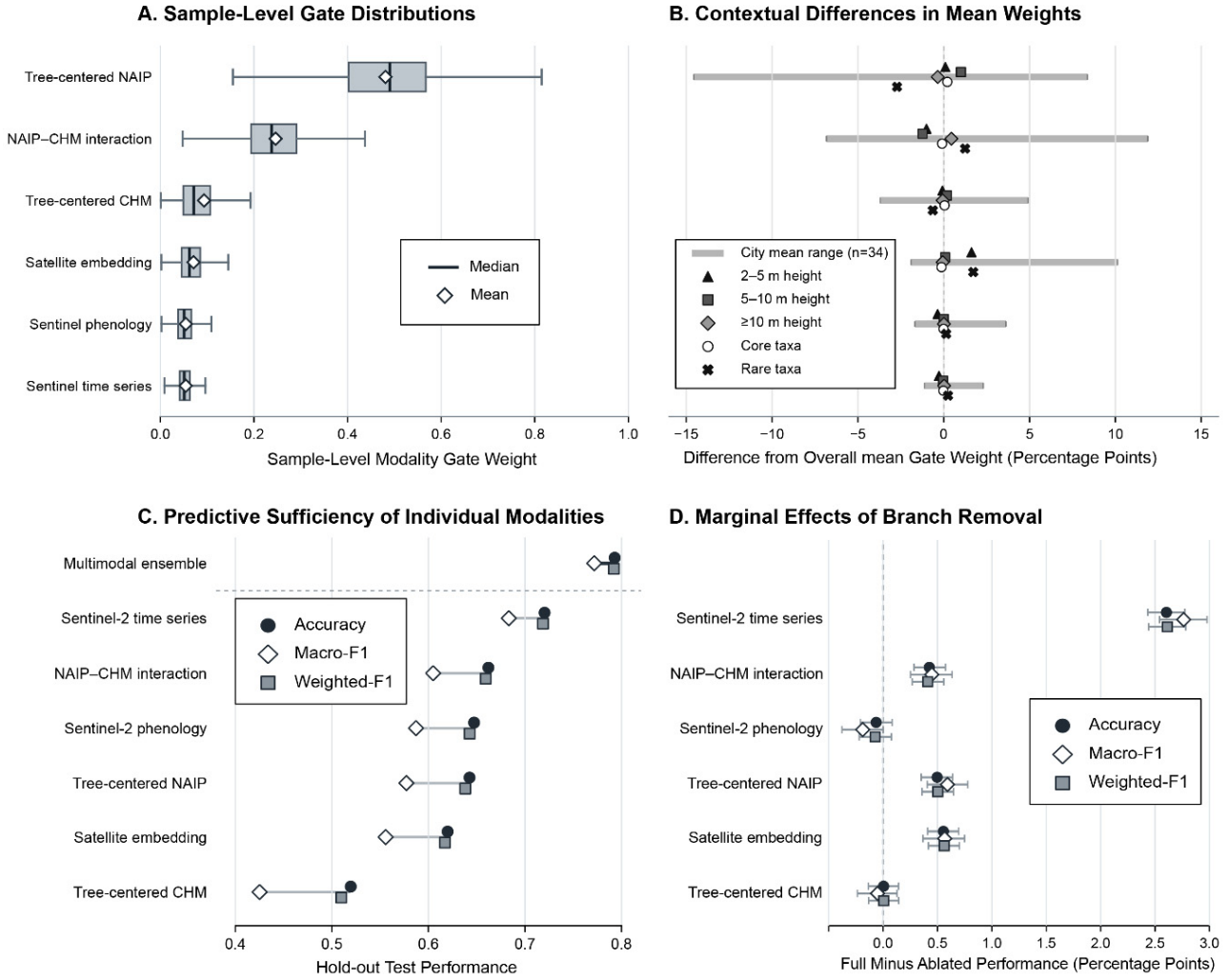


Figure 7. Modality use, predictive sufficiency, and marginal contribution in the final multimodal classifier. (**A**) Distributions of sample-specific gate weights across independent test records. (**B**) Differences in mean gate weights across cities, vegetation-height strata, and core versus rare taxa, relative to the overall mean for each modality. Gray bars show the range of city-level means. (**C**) Held-out test accuracy, macro-F1, and weighted-F1 for models trained using individual modalities, with the final multimodal ensemble included as a reference. (**D**) Changes in performance between the complete model and leave-one-branch-out models; positive values indicate poorer performance after removing the branch, and error bars show paired-bootstrap 95% confidence intervals based on 2000 replicates. Panels (**C**,**D**) report performance after probability-sum collapse from 11 intermediate groups to six reported groups. Gate weights indicate model allocation among available representations and should not be interpreted as causal feature importance. Abbreviations: NAIP—National Agricultural Imagery Program, CHM—Canopy Height Model.

Gate weights did not directly correspond to predictive importance. Despite tying for the lowest mean gate weight, the raw Sentinel-2–PRISM branch was the strongest single-modality model, achieving an accuracy of 0.720 and macro-F1 of 0.683 after collapse to six groups. Its removal also produced the largest ablation effect, reducing accuracy, macro-F1, and weighted-F1 by approximately 2.6–2.8 percentage points. Removing tree-centered NAIP, the NAIP–CHM interaction, or satellite embeddings produced smaller but statistically significant reductions of approximately 0.4–0.6 percentage points. Effects from removing tree-centered CHM or Sentinel-2 phenology were not statistically significant. Together, these results show that the classifier integrated complementary spatial and temporal information and that gate weights should not be interpreted as causal measures of modality importance.

3.6. Geographic Variation and Sensitivity

The default model achieved an overall independent-test accuracy of 0.793, macro-F1 of 0.772, and weighted-F1 of 0.792. Performance nevertheless varied substantially among the 22 cities with at least 1000 quality-screened test records (Figure 8, Table S14). City-level accuracy ranged from 0.728 in Baltimore to 0.955 in Cape Coral, whereas macro-F1 ranged from 0.443 in Denver to 0.864 in Sioux Falls. Mean city-level accuracy was 0.811, but mean macro-F1 was 0.638, indicating that high accuracy sometimes reflected strong prediction of locally prevalent groups rather than consistently balanced performance across groups.

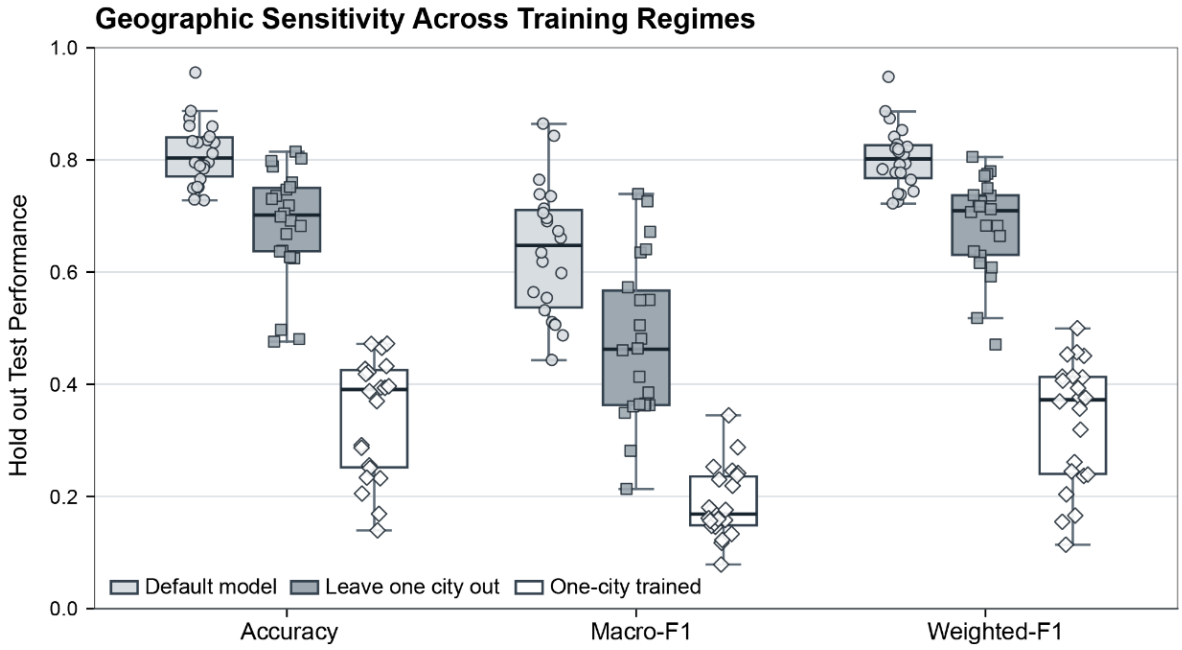


Figure 8. Geographic Sensitivity of Model Performance Across Training Regimes. Independent test accuracy, macro-F1, and weighted-F1 are shown for the default model trained with all available cities, leave-one-city-out models trained without the target city, and models trained using a single source city. Each point represents performance for one of 22 target cities; for the single-city regime, values were averaged across the seven eligible source-city models for each target city. Boxes span the interquartile range, center lines indicate medians, and whiskers extend to 1.5 times the interquartile range.

Withholding the target city from training substantially reduced performance. Mean accuracy and macro-F1 declined from 0.812 and 0.640 under the primary split to 0.685 and 0.478 under leave-one-city-out (LOCO) evaluation; corresponding LOCO medians were 0.702 and 0.462. LOCO accuracy was highest in Garden Grove (0.815) and lowest in San Jose (0.476), whereas macro-F1 was highest in Aurora (0.739) and lowest in Cape Coral (0.213). Models trained using only one city performed considerably worse, averaging 0.343 accuracy

and 0.188 macro-F1 across target cities, with corresponding medians of 0.390 and 0.168. Averaged across the seven source-city models, Seattle had the highest target-city accuracy (0.472) and Las Vegas the lowest (0.139); Aurora had the highest macro-F1 (0.344), whereas Cape Coral had the lowest (0.078). Among the source cities, the Portland-trained model produced the highest mean cross-city accuracy (0.391), the Denver-trained model produced the highest mean macro-F1 (0.227), and the Rancho Cucamonga-trained model produced the lowest values for both metrics (accuracy = 0.247; macro-F1 = 0.162). These results demonstrate that multisite training substantially improved geographic transfer and that models trained from a single municipal inventory generally lacked sufficient geographic and group-level diversity for national application.

Figure 9 shows correct and incorrect LOCO predictions for Rancho Cucamonga, California. Rancho Cucamonga was selected because its accuracy (0.637) and macro-F1 (0.505) were jointly close to the corresponding city medians and all six collapsed groups were represented. Overall errors exhibited weak but statistically significant spatial autocorrelation (Moran's $I = 0.079$, $p = 0.001$). Autocorrelation was strongest for Groups 3 and 4 (Moran's $I = 0.171$ and 0.136, respectively; both $p = 0.001$), followed by Groups 1 and 5 (0.079, $p = 0.001$; and 0.073, $p = 0.006$). Thus, although errors were spatially structured within some groups, their overall clustering was limited.

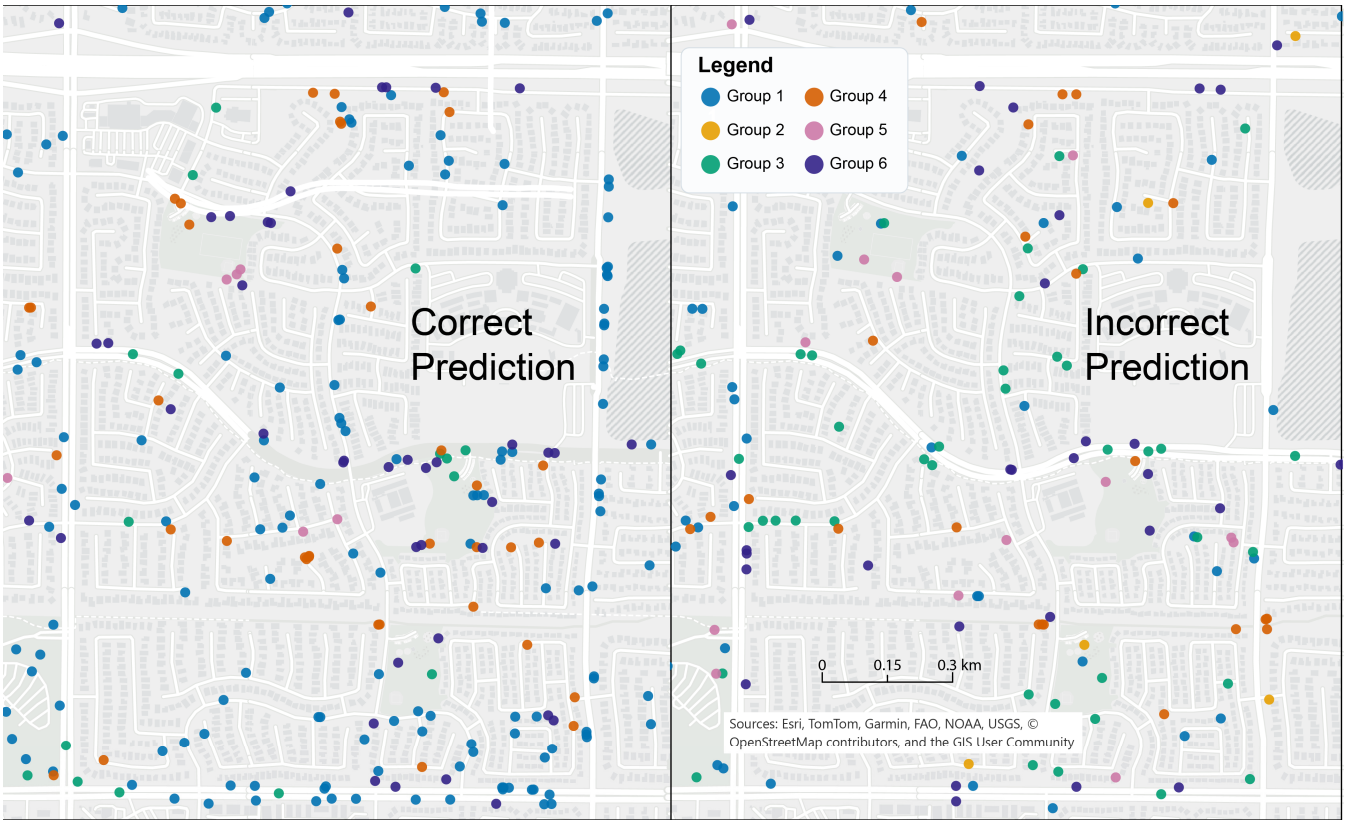


Figure 9. Leave-One-City-Out Model Predictions in Rancho Cucamonga, California. Predictions are from the fixed held-out test partition records. (**Left**): model predictions that matched group labels. (**Right**): model predictions that did not match group labels. Errors exhibit weak spatial correlation overall, with the greatest correlation for groups 3 and 4.

4. Discussion

We developed a multimodal deep-learning framework integrating municipal tree inventories with high-resolution aerial imagery, LiDAR-derived canopy height, Sentinel-2 phenology, satellite embeddings, and weather data to classify individual urban trees

across U.S. cities. Among 1.85 million eligible records, the final pruned ensemble predicted 11 intermediate groups that were subsequently collapsed into six reported groups. On the independent test set, the collapsed model achieved an accuracy of 0.793, macro-F1 of 0.772, and weighted-F1 of 0.792, with group-specific F1 scores ranging from 0.705 to 0.867. Performance was stronger for records with LiDAR-derived vegetation height ≥ 2 m and varied among groups and cities. Although the gating network assigned its greatest mean weights to NAIP-based inputs, ablation results showed that high-resolution spatial and coarser temporal inputs contributed complementary information. Geographic transfer declined when target cities were withheld, but remained substantially stronger than transfer from models trained on individual cities. Together, these findings demonstrate the feasibility—and current limits—of broad-scale urban-tree classification at a taxonomic resolution supported by publicly available remote-sensing data.

4.1. Data Availability, Quality, and Standardization

The growing availability of public imagery, LiDAR, satellite products, and accessible processing tools makes broad-scale urban-tree modeling increasingly feasible without costly data purchases or proprietary remote-sensing software. However, the quality, spatial resolution, acquisition timing, and completeness of these inputs varied substantially among cities.

Inventory quality was a major source of attrition. Inventories from cities in which more than 50% of georeferenced records shared coordinates accounted for approximately one-third of the original records and were excluded because their locations could not reliably identify individual trees. Additional records were removed because they had duplicated coordinates, invalid taxonomic labels, or could not be matched to detected crowns. LiDAR availability also varied markedly across cities, with some locations having only older datasets available through our data sources, and others having incomplete spatial coverage. Nearly half of Atlanta's inventory records were excluded because fewer than 50% of the 1 m cells within 20 m of the matched crown contained LiDAR observations.

NAIP products differed in acquisition year, spatial resolution, and file structure. Older products provided visible and near-infrared information in separate three-band files, whereas newer products combined the required bands in a single multiband product. Spatial resolution ranged from 0.08 to 0.6 m, requiring resampling to a common grid and consequently discarding some detail from the highest-resolution imagery. LiDAR point density also varied substantially, requiring aggregation to 1 m canopy-height products. These differences illustrate the tradeoff between the broad availability of public remote-sensing data and the extensive preprocessing needed to make observations comparable across locations and acquisition periods.

Residual geographic domain shifts likely remained after spatial standardization because NAIP resolution and acquisition year, LiDAR acquisition year and density, inventory practices, and urban context varied among cities. Temporal-mismatch sensitivity analyses identified metric-specific differences, but acquisition properties were strongly confounded with city and could not be interpreted independently. The leave-one-city-out analysis instead captured their combined influence on transfer to unseen cities.

Quality-control filtering was not random and therefore limits representativeness. Retained observations disproportionately came from cities with unique tree coordinates, sufficiently recent and complete LiDAR, available multimodal imagery, and standardized taxonomic labels. Within cities, filtering favored trees that could be matched to detected crowns and had adequate vegetation and LiDAR support. Performance estimates therefore apply most directly to data-rich municipal inventories and may not transfer equally to

cities, taxa, or small trees disproportionately removed during preprocessing. City-specific retention is reported in Table S1 and the complete screening workflow in Figure S1.

4.2. Taxonomic Groupings

The groups used in this study were derived empirically to maximize separation in the discriminator embedding space. This approach targeted taxonomic groupings distinguishable from the available remote-sensing inputs and may approximate the level of taxonomic resolution attainable with these data. However, remote-sensing separability does not establish shared ecological function. The groups should therefore not be interpreted as functional types or applied directly to ecosystem-service estimates, such as air-pollution mitigation or allergenic-pollen exposure, without independent trait-based validation.

Future work could construct groups using both empirical separability and ecological information. Where prior knowledge supports assigning a boundary taxon to an adjacent group, the partition could be revised and the classifier retrained or recalibrated. For example, the groups identified here could already provide spatial information on broad urban tree assemblages relevant to pollen production, while future classifications aligned with pollen-relevant genera or families could enable more taxon-specific predictions. These classifications could be integrated with flowering phenology, meteorology, and pollen monitoring data to improve estimates of within-city pollen concentrations. Any such modification should be evaluated using independent data to determine whether greater ecological coherence can be achieved without substantially reducing predictive performance.

The 50-record lower bound balanced taxonomic coverage against assignment reliability. Within the fixed discriminator space, rarefaction indicated that taxon centroids and nearest-group assignments generally stabilized by 50 records (Figure S14). Directly lowering the discriminator-training threshold from 500 to 50 records would have increased the number of labels from 280 to 579 while adding only 57,500 records (3.2%), more than doubling the classification problem with comparatively little additional support per label (Figure S15). We therefore trained the discriminator using 280 core taxa and assigned the additional 299 taxa to fixed prototypes afterward. Taxa represented by fewer than 50 records remained excluded, limiting applicability to the least-represented taxa.

4.3. Model Performance and Group Balance

Overall performance was encouraging for a model spanning the conterminous United States. Performance was substantially higher for records with maximum LiDAR-derived vegetation height ≥ 2 m, suggesting that larger, more mature trees were easier to classify than shorter or juvenile trees. However, reduced performance below 2 m may also reflect input-data errors. Crown detectors can identify tall shrubs as trees, while municipal inventories may contain positional, taxonomic, or temporal errors that produce incorrect crown–inventory matches. This label-alignment problem affects individual-tree classification generally and cannot be eliminated solely through model training. The height-stratified results therefore reflect both biological differences and variation in data quality.

Collapsing the 11 intermediate groups to six reported groups produced a more balanced label space than the earlier partition. Groups 2 and 3 each contained approximately 35% of quality-screened test records, Group 1 contained approximately 14%, and Groups 4–6 each contained approximately 5%. Group-level F1 was moderately but nonsignificantly associated with test support (Pearson's $r = 0.44$, $p = 0.38$). The highest F1 occurred for Group 1 (0.867), rather than either of the two largest groups, while the similarly sized Groups 4–6 had F1 scores of 0.709, 0.787, and 0.705. Support alone therefore did not explain group-level performance. Instead, the reciprocal confusion between Groups 2 and 3 and the

prototype-margin analysis indicate that overlap in remote-sensing characteristics remained an important source of error.

Many species-classification studies restrict training and evaluation to a small number of common species [10–15], leaving performance on uncommon or unmodeled taxa unmeasured. Our framework included 579 taxa by training the discriminator on 280 core taxa and assigning an additional 299 rare taxa to fixed group prototypes. This substantially expanded taxonomic coverage, although taxa with fewer than 50 records remained excluded and group imbalance was not eliminated.

The final model was trained using one fixed random seed, so its evaluation does not isolate variation caused by stochastic optimization. The five-checkpoint ensemble represents temporal averaging within that training run rather than independent repetitions. Although the independently trained geographic sensitivity models characterize variation caused by changing city composition, repeated-seed experiments would still be required to distinguish optimization variability from geographic sensitivity.

4.4. Modality Contributions and Error Propagation

Gate weights were concentrated on tree-centered NAIP and the NAIP–CHM interaction, indicating that high-resolution spatial information dominated the fused representation. However, gate weights did not measure marginal predictive importance. The joint raw Sentinel-2–PRISM branch was the strongest single-branch model, and its removal caused the largest reductions in accuracy and macro-F1 (2.6 and 2.8 percentage points, respectively; Figure 7). The temporal sequence therefore provided information not fully captured by the high-resolution spatial inputs, although its design did not permit the contributions of Sentinel-2 and PRISM to be separated. Gate allocations also varied among observations and cities, but these patterns cannot establish why particular inputs were up- or downweighted.

Sentinel-2 measurements represented 10 m cells rather than individual crowns. Centerpoint assignment could therefore produce label–feature mismatch when cells contained multiple crowns, shrubs, lawns, or impervious surfaces. Lower performance under limited within-cell NAIP vegetation and tall-CHM coverage was consistent with mixed-pixel effects, although these associations could not separate contamination from crown size, surrounding vegetation, or geographic context. Performance also declined modestly with extensive temporal interpolation. Sentinel-2 should consequently be interpreted as providing a mixture of tree, local-canopy, and regional seasonal context rather than a pure individual-crown signal. Future work could evaluate crown-fraction unmixing, higher-resolution temporal imagery, or crown-size-dependent use of Sentinel-2 inputs.

Crown detection and inventory matching could further propagate omission, commission, and label–feature errors into training. Reduced sample weights limited the influence of low-height or distant matches but could not correct erroneous crown–inventory pairs. This is a broader limitation of individual-tree remote-sensing classification, which generally relies on imperfect crown delineation or geolocation to associate imagery with reference labels. Without independently validated crown–inventory links, the contribution of this error source could not be isolated.

4.5. Geographic Generalizability

Performance under the primary split varied substantially among cities. Accuracy ranged from 0.728 to 0.955, while macro-F1 ranged from 0.443 to 0.864. The divergence between these metrics indicates that high city-level accuracy sometimes reflected successful prediction of locally prevalent groups rather than balanced performance across all groups present. Local group composition should therefore be considered when interpreting geographic variation.

Because the primary record-level split allowed the same cities—and potentially nearby trees—to occur across partitions, it primarily measured prediction of new trees within represented geographic domains. Nearby observations could share Sentinel-2 cells, satellite embeddings, PRISM conditions, or overlapping NAIP and CHM context, making this evaluation optimistic for unseen cities. Consistent with this limitation, mean accuracy and macro-F1 declined from 0.812 and 0.640 under the primary split to 0.685 and 0.478 when target cities were withheld. Geographic transfer was therefore substantially more difficult than within-city interpolation.

Nevertheless, multisite training generalized much better than training on a single municipal inventory. Across target cities, the seven single-city-trained models averaged only 0.343 accuracy and 0.188 macro-F1. This comparison demonstrates the value of geographically diverse training data and represents a substantial improvement over the common practice of developing urban-tree classifiers from a single city [10–15]. The remaining LOCO performance loss indicates that additional geographic coverage would likely improve transfer further.

Differences in inventory practices, duplicate-coordinate prevalence, group composition, imagery acquisition timing, LiDAR age and density, climate, and urban form could all contribute to intercity variation. Many of these properties were shared within cities and therefore could not be separated statistically from geographic location. More standardized municipal inventory protocols and more balanced within-city sampling could improve both model development and geographic comparability. Future work should also examine variation at finer spatial scales, including local climate zones.

The conterminous United States provides a substantially more heterogeneous modeling domain than the smaller study areas used in many previous classification studies. A national model, however, may sacrifice distinctions recoverable within regions containing fewer taxa and more homogeneous environmental and cultural conditions. A useful compromise may be a set of models defined for a limited number of climatic or culturally coherent regions. Alternative regionalizations could be proposed using domain knowledge and then compared empirically using geographically independent validation.

5. Conclusions

Currently available public remote-sensing data do not provide sufficient information to classify hundreds of urban tree species reliably across broad geographic extents. However, our results demonstrate that multimodal observations can classify individual trees into broader groups defined by similarities in remotely sensed canopy appearance, structure, phenology, and environmental context. High-resolution NAIP and LiDAR inputs provided the dominant spatial information, while Sentinel-2 and PRISM contributed complementary temporal context despite their coarser resolution. Performance declined for unseen cities but remained substantially stronger with multisite training than with models trained on individual municipal inventories. The reported outputs are six remote-sensing-derived taxonomic groups and should not be interpreted as species-level identifications or independently validated ecological functional types.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs18213696/s1, Table S1, City-Level Tree Inventory Composition and Filtering Outcomes; Figure S1, Sequential Dataset Attrition and Use of Fixed Analytical Partitions; Figure S2, Distribution of Duplicate-Coordinate Records across the 60 Original Municipal Tree Inventories; Table S2, NAIP Imagery Characteristics by City; Figure S3, Difference between NAIP and Tree Inventory Acquisition Years; Table S3, LiDAR Characteristics by City; Figure S4, Temporal Difference between Tree Inventory and LiDAR Acquisition Years; Table S4, Daily PRISM Meteorological Variables Aligned with the Sentinel-2 Observation Sequence; Figure S5, Tree-Crown Detection and

Inventory-Matching Workflow; Figure S6, Tree-Crown Matching-Distance Distribution; Figure S7, Tree-Centered NAIP Image Channels; Figure S8, Tree-Centered LiDAR Canopy Products; Table S5, Structural Summary Variables Derived from NAIP and CHM Chips; Table S6, Raw Sentinel-2 Sequence Variables; Figure S9, Sentinel-2 NDVI Observations and Interpolation; Figure S10, Distribution of Sentinel-2 Interpolation-Gap Durations; Figure S11, Sentinel-2 Interpolation-Gap Durations by City; Table S7, Derived Sentinel-2 Phenology Variables; Table S8, Configuration of the Remote-Sensing-Derived Taxonomic Discriminator Model; Table S9, Configuration of the Stage-Two Classification Model; Table S10, Validation Performance and Group Balance of Candidate Probability-Sum Collapses of the 11 Intermediate Groups; Table S11, Leave-One-Branch-Out Modality-Pruning Analysis; Section S1, Core Taxa Assigned to the 11 Intermediate Classification Groups; Section S2, Rare Taxa Assigned to the 11 Intermediate Classification Groups; Table S12, Independent-Test Performance of the Taxon-Discrimination Model and Its Derived Group Classifications; Table S13, Held-Out Test Performance of the Final Pruned Stage-Two Classifier across the 11 Intermediate Groups; Figure S12, Confusion Matrix for the 11 Intermediate Groups Predicted by the Final Pruned Stage-Two Classifier; Figure S13, Relationship between Prototype Margin and Taxon-Level Misclassification Error for the 11 Intermediate Remote-Sensing-Derived Taxonomic Groups; Table S14, Geographic Generalization of the Final Pruned Classifier under Primary, Leave-One-City-Out, and Single-City Training Scenarios; Figure S14, Sensitivity of Taxon-Centroid Estimation to Record Support; Figure S15, Tradeoff between Taxonomic Coverage and Record Support per Label.

Author Contributions: A.L. contributed to study design, implementation, analysis, manuscript writing and editing. M.B. contributed to study design, analysis, and manuscript editing. P.H. contributed to study design, analysis, and manuscript editing. All authors have read and agreed to the published version of the manuscript.

Funding: Research reported in this publication was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award R01HL150119 and R01HL178727. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability Statement: Data sources used for this study are publicly available and listed in Table 1. Computer code, training datasets, and models created from this study are available at the GitHub repository https://github.com/larkinandy/remoteUrbanTreeId, accessed on 20 October 2026.

Acknowledgments: OpenAI Codex 5.6 Sol Light was used to create most of the computer scripts used in this study, including scripts to download and preprocess datasets, train and evaluate models, and calculate summary statistics. All generated code is available at the GitHub repository referenced in the Data Availability Statement. During the preparation of this manuscript, the authors also used Codex for language editing and improvements to clarity and readability, and to generate Tables S1–S9, Sections S1 and S2, and Figures S1, S14, and S15. The authors reviewed and edited all output and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rahman, M.A.; Stratopoulos, L.M.F.; Moser-Reischl, A.; Zölch, T.; Häberle, K.-H.; Rötzer, T.; Pretzsch, H.; Pauleit, S. Traits of Trees for Cooling Urban Heat Islands: A Meta-Analysis. *Build. Environ.* **2020**, *170*, 106606. [CrossRef]
2. Pregitzer, C.C.; Hanna, C.; Charlop-Powers, S.; Bradford, M.A. Estimating Carbon Storage in Urban Forests of New York City. *Urban Ecosyst.* **2022**, *25*, 617–631. [CrossRef]
3. Carlyle-Moses, D.E.; Livesley, S.; Baptista, M.D.; Thom, J.; Szota, C. Urban Trees as Green Infrastructure for Stormwater Mitigation and Use. In *Forest-Water Interactions*; Levia, D.F., Carlyle-Moses, D.E., Iida, S., Michalzik, B., Nanko, K., Tischer, A., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 397–432, ISBN 978-3-030-26086-6.
4. Rathmann, J.; Beck, C.; Flutura, S.; Seiderer, A.; Aslan, I.; André, E. Towards Quantifying Forest Recreation: Exploring Outdoor Thermal Physiology and Human Well-Being along Exemplary Pathways in a Central European Urban Forest (Augsburg, SE-Germany). *Urban For. Urban Green.* **2020**, *49*, 126622. [CrossRef]
5. Zhang, D.; Yao, M.; Chen, Y.; Liu, Y. The Role of Urban Vegetation in Mitigating Fire Risk Under Climate Change: A Review. *Sustainability* **2025**, *17*, 2680. [CrossRef]

6. Tegart, L.J.; Johnston, F.H.; Borchers Arriagada, N.; Workman, A.; Dickinson, J.L.; Green, B.J.; Jones, P.J. 'Pollen Potency': The Relationship between Atmospheric Pollen Counts and Allergen Exposure. *Aerobiologia* **2021**, *37*, 825–841. [CrossRef]
7. Roman, L.A.; Conway, T.M.; Eisenman, T.S.; Koeser, A.K.; Ordóñez Barona, C.; Locke, D.H.; Jenerette, G.D.; Östberg, J.; Vogt, J. Beyond 'Trees Are Good': Disservices, Management Costs, and Tradeoffs in Urban Forestry. *Ambio* **2021**, *50*, 615–630. [CrossRef] [PubMed]
8. Austin, K.G.; Baker, J.S.; Sohngen, B.L.; Wade, C.M.; Daigneault, A.; Ohrel, S.B.; Ragnauth, S.; Bean, A. The Economic Costs of Planting, Preserving, and Managing the World's Forests to Mitigate Climate Change. *Nat. Commun.* **2020**, *11*, 5946. [CrossRef] [PubMed]
9. McCoy, D.E.; Goulet-Scott, B.; Meng, W.; Atahan, B.F.; Kiros, H.; Nishino, M.; Kartesz, J. Species Clustering, Climate Effects, and Introduced Species in 5 Million City Trees across 63 US Cities. *eLife* **2022**, *11*, e77891. [CrossRef] [PubMed]
10. Zhang, Z.; Kazakova, A.; Moskal, L.M.; Styers, D.M. Object-Based Tree Species Classification in Urban Ecosystems Using LiDAR and Hyperspectral Data. *Forests* **2016**, *7*, 122. [CrossRef]
11. Hartling, S.; Sagan, V.; Sidike, P.; Maimaitijiang, M.; Carron, J. Urban Tree Species Classification Using a WorldView-2/3 and LiDAR Data Fusion Approach and Deep Learning. *Sensors* **2019**, *19*, 1284. [CrossRef] [PubMed]
12. Li, H.; Hu, B.; Li, Q.; Jing, L. CNN-Based Individual Tree Species Classification Using High-Resolution Satellite Imagery and Airborne LiDAR Data. *Forests* **2021**, *12*, 1697. [CrossRef]
13. Fang, F.; McNeil, B.E.; Warner, T.A.; Maxwell, A.E.; Dahle, G.A.; Eutsler, E.; Li, J. Discriminating Tree Species at Different Taxonomic Levels Using Multi-Temporal WorldView-3 Imagery in Washington D.C., USA. *Remote Sens. Environ.* **2020**, *246*, 111811. [CrossRef]
14. Neyns, R.; Efthymiadis, K.; Libin, P.; Canters, F. Fusion of Multi-Temporal PlanetScope Data and Very High-Resolution Aerial Imagery for Urban Tree Species Mapping. *Urban For. Urban Green.* **2024**, *99*, 128410. [CrossRef]
15. Pu, R.; Landry, S. Mapping Urban Tree Species by Integrating Multi-Seasonal High Resolution Pléiades Satellite Imagery with Airborne LiDAR Data. *Urban For. Urban Green.* **2020**, *53*, 126675. [CrossRef]
16. Jiang, Y.; Li, X.; Peng, L.; Li, C.; Song, T. Assessing the Potential of Multi-Seasonal Sentinel-2 Satellite Imagery Combined with Airborne LiDAR for Urban Tree Species Identification. *Sci. Rep.* **2025**, *15*, 25107. [CrossRef] [PubMed]
17. Marks, C.O.; Nislow, K.H.; Magilligan, F.J. Quantifying Flooding Regime in Floodplain Forests to Guide River Restoration. *Elem. Sci. Anthr.* **2014**, *2*, 000031. [CrossRef]
18. Ventura, J.; Pawlak, C.; Honsberger, M.; Gonsalves, C.; Rice, J.; Love, N.L.R.; Han, S.; Nguyen, V.; Sugano, K.; Doremus, J.; et al. Individual Tree Detection in Large-Scale Urban Environments Using High-Resolution Multispectral Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *130*, 103848. [CrossRef]
19. Earth Resources Observation and Science (EROS) Center. National Agriculture Imagery Program (NAIP). 2003. Available online: https://data.usgs.gov/datacatalog/data/USGS:EROS5e83a340bf820c39 (accessed on 10 August 2026).
20. U.S. Geological Survey. *The National Map—New Data Delivery Homepage, Advanced Viewer, Lidar Visualization*; Fact Sheet 2019–3032; US. Geological Survey, 2019. Available online: https://pubs.usgs.gov/publication/fs20193032 (accessed on 10 August 2026).
21. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
22. Brown, C.F.; Kazmierski, M.R.; Pasquarella, V.J.; Rucklidge, W.J.; Samsikova, M.; Zhang, C.; Shelhamer, E.; Lahera, E.; Wiles, O.; Ilyushchenko, S.; et al. AlphaEarth Foundations: An Embedding Field Model for Accurate and Efficient Global Mapping from Sparse Label Data. *arXiv* **2025**, arXiv:2507.22291.
23. Di Luzio, M.; Johnson, G.L.; Daly, C.; Eischeid, J.K.; Arnold, J.G. Constructing Retrospective Gridded Daily Precipitation and Temperature Datasets for the Conterminous United States. *J. Appl. Meteorol. Climatol.* **2008**, *47*, 475–497. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.